Methods in functional data analysis and functional genomics

Daniel Backenroth

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
under the Executive Committee of the
Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Methods in functional data analysis and functional genomics

Daniel Backenroth

This thesis has two overall themes, both of which involve the word functional, albeit in different contexts. The theme that motivates two of the chapters is the development of methods that enable a deeper understanding of the variability of functional data. The theme of the final chapter is the development of methods that enable a deeper understanding of the landscape of functionality across the human genome in different human tissues.

The first chapter of this thesis provides a framework for quantifying the variability of functional data and for analyzing the factors that affect this variability. We extend functional principal components analysis by modeling the variance of principal component scores. We pose a Bayesian model, which we estimate using variational Bayes methods. We illustrate our model with an application to a kinematic dataset of two-dimensional planar reaching motions by healthy subjects, showing the effect of learning on motion variability.

The second chapter of this thesis provides an alternative method for decomposing functional data that follows a Poisson distribution. Classical methods pose a latent Gaussian process that is then linked to the observed data via a logarithmic link function. We pose an alternative model that draws on ideas from non-negative matrix factorization, in which we constrain both scores and spline coefficient vectors for the functional prototypes to be non-negative. We impose smoothness on the functional prototypes. We estimate our model using the method of alternating minimization. We illustrate our model with an application to a dataset of accelerometer readings from elderly healthy Americans.

The third chapter of this thesis focuses on functional genomics, rather than functional data analysis. Here we pose a method for unsupervised clustering of functional genomics data. Our method is non-parametric, allowing for flexible modeling of the functional ge-

nomics data without binarization. We estimate our model using variational Bayes methods, and illustrate it by calculating genome-wide functional scores (based on a partition of our clusters into functional and non-functional clusters) for 127 different human tissues. We show that these genome-wide and tissue-specific functional scores provide state-of-the-art functional prediction.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to Jeff Goldsmith, my advisor at the Mailman School of Public Health, for his support, guidance and encouragement throughout my graduate studies. He is a font of unfailingly good advice, and I will always be grateful for his ever-positive attitude.

I am grateful also to Iuliana Ionita-Laza for teaching me the fundamentals of statistical genetics and functional genomics, and for giving me the opportunity to work on interesting projects.

I am grateful to Ying Wei and Carol Friedman for giving me the opportunity to work with electronic health records data and for their guidance and support.

I am grateful to Todd Ogden and Tomoko Kitago, members of my doctoral committee, for their valuable thoughts and suggestions that have helped me to improve my thesis.

I am grateful to Taki Shinohara, Herbert Chase, Shai Carmi and David Zeevi for their support and guidance.

I am grateful to Yufeng Shen for having confidence in me and bringing me to Columbia, and for teaching me the fundamentals of bioinformatics and of the scientific process.

I am grateful to Shing Lee for her wise supervision and for welcoming me to the Mailman School.

I owe many thanks to Katy Hardy for her invaluable help.

Thank you to Lyudmila Ena for help with clinical data at the Department of Biomedical Informatics and to Justine Herrera for help navigating my way to a degree. Thank you to Olivier Couronne, Wei-Yann Tsai, Bin Cheng, Ken Cheung, Codruta Chiuzan, DuBois Bowman, my colleagues Samreen Zafer, Angad Singh, Jimmy Duong, Vivian Zhang, Xiaoyun Sun, and Boris Grinshpun, and my fellow students Ming Sun, Eun Jeung Oh, Yakuan Chen and Julia Wrobel, for their help in my research and work. Thank you to the Department of Biostatistics for its financial support for my studies.

Thank you also to Jennifer Schrack, Michelle Harran, Juan Cortes, John W. Krakauer, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Pethukova, Ekta Khurana, Angela Christiano and Joseph D. Buxbaum.

I give thanks to my parents for supporting me during my studies at Columbia, and for their love, and to my brother Ariel for his love and his help with my programming questions. I give thanks to my children, Isaac, Jonah, and Eli, for their love and for making every moment (well, almost every moment) with them so fun.

My most important thanks go to Paola, without whom I could never have even started, let alone completed, this project. As I write this on our twelfth anniversary, I thank you for always being there for me, for all the fun times we have had together, for your patience with me when my head is stuck in my work, for your love, for everything!

To Paola

Chapter 1

Introduction

The first part of this thesis focuses on functional data analysis, and specifically on functional principal component analysis and related methods.

The first chapter in the first part of this thesis develops a framework for quantifying the variability of functional data and for analyzing the factors that affect this variability. We start with functional principal components analysis, in which a basis for a set of functional observations is selected that is in some sense optimal for capturing the variability of those observations. Classically, the first FPC is selected to explain the largest possible amount of variance in the data, subject to some smoothness penalty. The second FPC is selected to be orthogonal to the first FPC, and to explain the largest possible amount of remaining variance in the data. And so on. In this framework, scores for each FPC are assumed to come from a distribution (usually Gaussian) with constant variance. We extend this model by allowing scores for each curve to have a different Gaussian distribution, whose variance depends on covariate and subject-specific factors. We fit our model using variational Bayes methods. In our application, our extended model enables us to quantify the effects of learning on motion variability, using a kinematic dataset of two-dimensional planar reaching motions by healthy subjects.

The second chapter in the first part of this thesis shifts focus from functions with Gaussian noise to functions that represent observations of a Poisson process. Classical methods for the analysis of this data pose a latent Gaussian process that is then linked to the observed data via a logarithmic link function. We pose an alternative model that draws on

ideas from non-negative matrix factorization, in which we constrain both scores and spline coefficient vectors for the functional prototypes to be non-negative. We impose smoothness on the functional prototypes. We estimate our model using the method of alternating minimization. We illustrate our model with an application to a dataset of accelerometer readings from elderly healthy Americans.

The second part of this thesis focuses on the use of statistical methods in functional genomics. A significant open problem in functional genomics is understanding the function of non-coding DNA, which comprises the vast majority of the human genome. Our contribution to this area of research is the development of a statistical method that can predict whether a given region of the genome is functional in a tissue-specific manner. This method makes use of a recent project, the Roadmap project, that generated tissue-specific histone mark binding and DNase hypersensitivity maps across the entire genome in more than a hundred tissues. Our method clusters this data probabilistically, non-parametrically modeling the distribution of histone mark binding and DNase measurements. Our method therefore differs from other existing methods that binarize this data, therefore losing some of the information in the assays. We estimate our model using variational Bayes methods, and illustrate it by calculating genome-wide functional scores (based on a partition of our clusters into functional and non-functional clusters) for 127 different human tissues. We show that these genome-wide and tissue-specific functional scores provide state-of-the-art functional prediction.

Part I

Methods in functional data analysis

Chapter 2

Modeling motor learning using heteroskedastic functional principal components analysis

2.1 Scientific motivation

2.1.1 Motor learning

Recent work in motor learning has suggested that change in motion variability is an important component of improvement in motor skill. It has been suggested that when a motor task is learned, variance is reduced along dimensions relevant to the successful accomplishment of the task, although it may increase in other dimensions [Scholz and Schöner, 1999; Yarrow *et al.*, 2009]. Experimental work, moreover, has shown that learning-induced improvement of motion execution, measured through the trade-off between speed and accuracy, is accompanied by significant reductions in motion variability. In fact, these reductions in motion variability may be a more important feature of learning than changes in the average motion [Shmuelof *et al.*, 2012]. These results have typically been based on assessments of variability at a few time points, e.g., at the end of the motion, although high-frequency laboratory recordings of complete motions are often available.

In this chapter we develop a modeling framework that can be used to quantify motion

variability based on dense recordings of fingertip position throughout motion execution. This framework can be used to explore many aspects of motor skill and learning: differences in baseline skill among healthy subjects, effects of repetition and training to modulate variability over time, or the effect of baseline stroke severity on motion variance and recovery [Krakauer, 2006]. By taking full advantage of high-frequency laboratory recordings, we shift focus from particular time points to complete curves. Our approach allows us to model the variability of these curves as they depend on covariates, like the hand used or the repetition number, as well as the estimation of random effects reflecting differences in baseline variability and learning rates among subjects.

Section 2.1.2 describes our motivating data in more detail, and Section 2.2 introduces our modeling framework. A review of relevant statistical work appears in Section 2.3. Details of our estimation approach are in Section 3.2. Simulations and the application to our motivating data appear in Sections 3.4 and 2.6, respectively, and we close with a discussion in Section 2.7.

2.1.2 Dataset

Our motivating data were gathered as part of a study of motor learning among healthy subjects. Kinematic data were acquired in a standard task used to measure control of reaching motions. In this task, subjects rest their forearm on an air-sled system to reduce effects of friction and gravity. The subjects are presented with a screen showing eight targets arranged in a circle around a starting point, and they reach with their arm to a target and back when it is illuminated on the screen. Subjects' motions are displayed on the screen, and they are rewarded with 10 points if they turn their hand around within the target, and 3 or 1 otherwise, depending on how far their hand is from the target at the point of return. Subjects are not rewarded for motions outside pre-specified velocity thresholds.

Our dataset consists of 9,481 motions by 26 right-handed subjects. After becoming familiarized with the experimental apparatus, each subject made 24 or 25 reaching motions to each of the 8 targets, in a semi-random order, with both the left and right hand. Motions that did not reach at least 30% of the distance to the target and motions with a direction more than 90° away from the target direction at the point of peak velocity were excluded

from the dataset, because of the likelihood that they were made to the wrong target or not attempted due to distraction. Motions made at speeds outside the range of interest, with peak velocity less than 0.04 or greater than 2.0 m/s, were also excluded. These exclusion rules and other similar rules have been used previously in similar kinematic experiments, and are designed to increase the specificity of these experiments for probing motor control mechanisms [Huang *et al.*, 2012; Tanaka *et al.*, 2009; Kitago *et al.*, 2015]. A small number of additional motions were removed from the dataset due to instrumentation and recording errors. The data we consider have not been previously reported.

For each motion, the X and Y position of the hand motion is recorded as a function of time from motion onset to the initiation of return to the starting point, resulting in bivariate functional observations denoted $[P_{ij}^X(t), P_{ij}^Y(t)]$ for subject i and motion j. In practice, observations are recorded not as functions but as discrete vectors. There is some variability in motion duration, which we remove for computational convenience by linearly registering each motion onto a common grid of length $D = 50$. The structure of the registered kinematic data is illustrated in Figure 2.1. The top and bottom rows show, respectively, the first and last right-hand motion made to each target by each subject. The reduction in motion variance after practice is clear.

Prior to our analyses, we rotate curves so that all motions extend to the target at 0°. This rotation preserves shape and scale, but improves interpretation. After rotation, motion along the X coordinate represents motion parallel to the line between origin and target, and motion along the Y coordinate represents motion perpendicular to this line. We build models for X and Y coordinate curves separately in our primary analysis. An alternative bivariate analysis appears in Appendix A.3.

2.2 Model for curve variance

We adopt a functional data approach to model position curves $P_{ij}(t)$. Here we omit the X and Y superscripts for notational simplicity. Our starting point is the functional principal component analysis (FPCA) model of Yao *et al.* [2005] with subject-specific means. In this

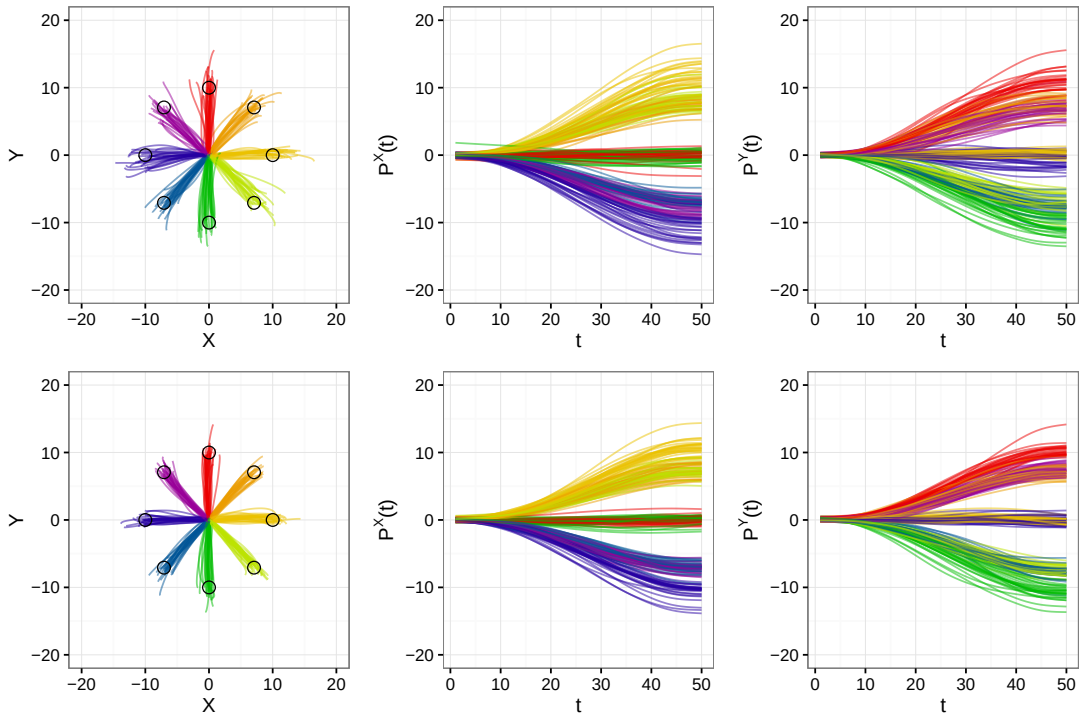

Figure 2.1: Observed kinematic data. The top row shows the first right-hand motion to each target for each subject; the bottom row shows the last motion. The left panel of each row shows observed reaching data in the X and Y plane. Targets are indicated with circles. The middle and right panels of each row show the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ curves, respectively.

model, it is assumed that each curve $P_{ij}(t)$ can be modeled as

$$\begin{aligned} P_{ij}(t) &= \mu_{ij}(t) + \delta_{ij}(t) \\ &= \mu_{ij}(t) + \sum_{k=1}^{\infty} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t). \end{aligned} \tag{2.1}$$

Here $\mu_{ij}(t)$ is the mean function for curve $P_{ij}(t)$, the deviation $\delta_{ij}(t)$ is modeled as a linear combination of eigenfunctions $\phi_k(t)$, the ξ_{ijk} are uncorrelated random variables with mean 0 and variances λ_k, where $\sum_k \lambda_k < \infty$ and $\lambda_1 \geq \lambda_2 \geq \cdots$, and $\epsilon_{ij}(t)$ is white noise. Here all the deviations $\delta_{ij}(t)$ are assumed to have the same distribution, that of a single underlying random process $\delta(t)$.

Model (2.1) is based on a truncation of the Karhunen-Loève representation of the random process $\delta(t)$. The Karhunen-Loève representation, in turn, arises from the spectral decomposition of the covariance of the random process $\delta(t)$ from Mercer's Theorem, from which one can obtain eigenfunctions $\phi_k(t)$ and eigenvalues λ_k.

The assumption of constant score variances λ_k in model (2.1) is inconsistent with our motivating data because it implies that the variability of the position curves $P_{ij}(t)$ is not covariate- or subject-dependent. However, motion variance can depend on the subject's baseline motor control and may change in response to training. Indeed, these changes in motion variance are precisely our interest.

In contrast to the preceding, we therefore assume that each random process $\delta_{ij}(t)$ has a potentially unique distribution, with a covariance operator that can be decomposed as

$$\mathrm{Cov}[\delta_{ij}(s), \delta_{ij}(t)] = \sum_{k=1}^{\infty} \lambda_{ijk}\phi_k(s)\phi_k(t),$$

so that the eigenvalues λ_{ijk}, but not the eigenfunctions, vary among the curves. We assume that deviations $\delta_{ij}(t)$ are uncorrelated across both i and j.

The model we pose for the $P_{ij}(t)$ is therefore

$$P_{ij}(t) = \mu_{ij}(t) + \sum_{k=1}^{K} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t), \tag{2.2}$$

where we have truncated the expansion in model (2.1) to K eigenfunctions, and into which we incorporate covariate and subject-dependent heteroskedasticity with the score variance

model

$$\lambda_{ijk} = \lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}} = \mathrm{Var}(\xi_{ijk}|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}) = \exp\left(\gamma_{0k} + \sum_{l=1}^{L^*}\gamma_{lk}x^*_{ijlk} + \sum_{m=1}^{M^*} g_{imk}z^*_{ijmk}\right) \tag{2.3}$$

where, as before, ξ_{ijk} is the kth score for the jth curve of the ith subject. In model (2.3), γ_{0k} is an intercept for the variance of the scores, γ_{lk} are fixed effects coefficients for covariates x^*_{ijlk}, $l = 1,\ldots,L^*$, and g_{imk} are random effects coefficients for covariates z^*_{ijmk}, $m = 1,\ldots,M^*$. The vector $\boldsymbol{g}_{ik}$ consists of the concatenation of the coefficients g_{imk}, and likewise for the vectors $\boldsymbol{x}^*_{ijk}$ and $\boldsymbol{z}^*_{ijk}$. Throughout, the subscript k indicates that models are used to describe the variance of scores associated with each basis function $\phi_k(t)$ separately. The covariates x^*_{ijlk} and z^*_{ijmk} in model (2.3) need not be the same across principal components. This model allows exploration of the dependence of motion variability on covariates, like progress through a training regimen, as well as of idiosyncratic subject-specific effects on variance through the incorporation of random intercepts and slopes.

Together, models (2.2) and (2.3) induce a subject- and covariate-dependent covariance structure for $\delta_{ij}(t)$:

$$\mathrm{Cov}[\delta_{ij}(s),\delta_{ij}(t)|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\phi_k,\boldsymbol{g}_{ik}] = \sum_{k=1}^{K}\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\phi_k(s)\phi_k(t).$$

In particular, the $\phi_k(t)$ are assumed to be eigenfunctions of a conditional covariance operator. Our proposal can be related to standard FPCA by considering covariate values random and marginalizing across the distribution of random effects and covariates using the law of total covariance:

$$\begin{aligned}\mathrm{Cov}[\delta_{ij}(s),\delta_{ij}(t)] &= E\left\{\mathrm{Cov}[\delta_{ij}(s),\delta_{ij}(t)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]\right\} + \\ &\quad \mathrm{Cov}\left\{E[\delta_{ij}(s)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]E[\delta_{ij}(t)|\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{g}]\right\} \\ &= \textstyle\sum_{k=1}^{K} E\left[\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\right]\phi_k(s)\phi_k(t).\end{aligned}$$

We assume that the basis functions $\phi_k(t)$ do not depend on covariate or subject effects, and are therefore unchanged by this marginalization. Scores ξ_{ijk} are marginally uncorrelated over k; this follows from the assumption that scores are uncorrelated in our conditional specification, and holds even if random effects $\boldsymbol{g}_{ik}$ are correlated over k. Lastly, the order

of marginal variances $E\left[\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}\right]$ may not correspond to the order of conditional variances $\lambda_{k|\boldsymbol{x}^*_{ijk},\boldsymbol{z}^*_{ijk},\boldsymbol{g}_{ik}}$ for some or even all values of the covariates and random effects coefficients.

In our approach, we assume that the scores ξ_{ijk} have mean zero. For this assumption to be valid, the mean $\mu_{ij}(t)$ in model (2.2) should be carefully modeled. To this end we use the well-studied multilevel function-on-scalar regression model [Guo, 2002; Di *et al.*, 2009; Morris and Carroll, 2006; Scheipl *et al.*, 2015],

$$\mu_{ij}(t) = \beta_0(t) + \sum_{l=1}^{L} x_{ijl}\beta_l(t) + \sum_{m=1}^{M} z_{ijm}b_{im}(t). \tag{2.4}$$

Here $\beta_0(t)$ is the functional intercept; x_{ijl} for $l \in 1, \ldots, L$ are scalar covariates associated with functional fixed effects with respect to the curve $P_{ij}(t)$; $\beta_l(t)$ is the functional fixed effect associated with the lth such covariate; z_{ijm} for $m \in 1, \ldots, M$ are scalar covariates associated with functional random effects with respect to the curve $P_{ij}(t)$; and $b_{im}(t)$ for $m \in 1, \ldots, M$ are functional random effects associated with the ith subject.

Keeping the basis functions constant across all subjects and motions, as in conventional FPCA, maintains the interpretability of the basis functions as the major patterns of variation across curves. Moreover, the covariate and subject-dependent score variances reflect the proportion of variation attributable to those patterns. To examine the appropriateness of this assumption for our data, we estimated basis functions for various subsets of motions using a traditional FPCA approach, after rotating observed data so that all motions extend to the target at 0°. As illustrated in Figure 2.2, the basis functions for motions made by both hands and at different stages of training are similar.

2.3 Prior work

FPCA has a long history in functional data analysis. It is commonly performed using a spectral decomposition of the sample covariance matrix of the observed functional data [Ramsay and Silverman, 2005; Yao *et al.*, 2005]. Most relevant to our current work are probabilistic and Bayesian approaches based on non-functional PCA methods [Tipping and Bishop, 1999; Bishop, 1999; Peng and Paul, 2009]. Rather than proceeding in stages, first

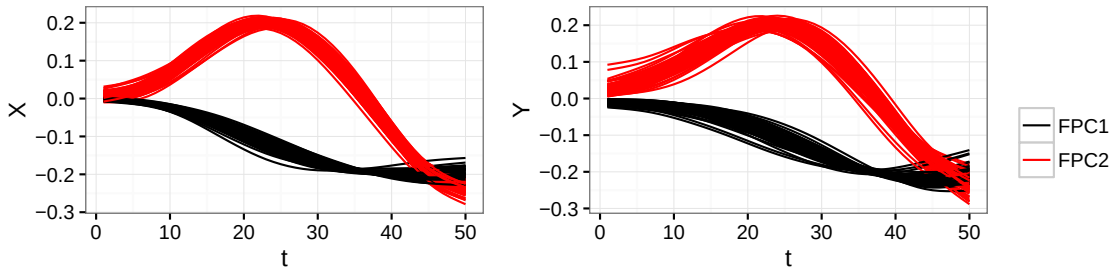

Figure 2.2: FPC basis functions estimated for various data subsets after rotating curves onto the positive X axis. The left panel shows the first and second FPC basis functions estimated for the X coordinate of motions to each target, for the left and right hand separately, and separately for motion numbers 1-6, 7-12, 13-18 and 19-24. The right panel shows the same for the Y coordinate.

by estimating basis functions and then, given these, estimating scores, such approaches estimate all parameters in model (2.1) jointly. James *et al.* [2000] focused on sparsely observed functional data and estimated parameters using an EM algorithm; van der Linde [2008] took a variational Bayes approach to estimation of a similar model. Goldsmith *et al.* [2015] considered both exponential-family functional data and multilevel curves, and estimated parameters using Hamiltonian Monte Carlo.

Some previous work has allowed for heteroskedasticity in FPCA. Chiou *et al.* [2003] developed a model which uses covariate-dependent scores to capture the covariate dependence of the mean of curves. In a manner that is constrained by the conditional mean structure of the curves, some covariate dependence of the variance of curves is also induced; the development of models for score variance was, however, not pursued. Here, by contrast, our interest is to use FPCA to model the effects of covariates on curve variance, independently of the mean structure. We are not using FPCA to model the mean; rather, the mean is modeled by the function-on-scalar regression model (2.4). Jiang and Wang [2010] introduce heteroskedasticity by allowing both the basis functions and the scores in an FPCA decomposition to depend on covariates. Briefly, rather than considering a bivariate covariance as the object to be decomposed, the authors pose a covariance surface that depends smoothly on a covariate. Aside from the challenge of incorporating more than a few covariates or

subject-specific effects, it is difficult to use this model to explore the effects of covariates on heteroskedasticity: covariates affect both the basis functions and the scores, making the interpretation of scores and score variances at different covariate levels unclear. Although it does not allow for covariate-dependent heteroskedasticity, the model of Huang *et al.* [2014] allows curves to belong to one of a few different clusters, each with its own FPCs and score variances.

In contrast to the existing literature, our model provides a general framework for understanding covariate and subject-dependent heteroskedasticity in FPCA. This allows the estimation of rich models with multiple covariates and random effects, while maintaining the familiar interpretation of basis functions, scores, and score variances.

Variational Bayes methods, which we use here to approximate Bayesian estimates of the parameters in models (2.2) and (2.3), are computationally efficient and typically yield accurate point estimates for model parameters, although they provide only an approximation to the complete posterior distribution and inference may suffer as a result [Ormerod and Wand, 2012; Jordan, 2004; Jordan *et al.*, 1999; Titterington, 2004]. These tools have previously been used in functional data analysis [van der Linde, 2008; Goldsmith *et al.*, 2011; McLean *et al.*, 2013]; in particular, Goldsmith and Kitago [2016] used variational Bayes methods in the estimation of model (2.4).

2.4 Methods

The main contribution of this chapter is the introduction of subject and covariate effects on score variances in model (2.3). Several estimation strategies can be used within this framework. Here we describe three possible approaches. Later, these will be compared in simulations.

2.4.1 Sequential estimation

Models (2.2) and (2.3) can be fit sequentially in the following way. First, the mean $\mu_{ij}(t)$ in model (2.2) is estimated through function-on-scalar regression under a working independence assumption of the errors; we use the function `pffr` in the `refund` package [Crainiceanu

et al., 2012] in `R`. Next, the residuals from the function-on-scalar regression are modeled using standard FPCA approaches to obtain estimates of principal components and marginal score variances; given these quantities, scores themselves can be estimated [Yao *et al.*, 2005]. For this step we use the function `fpca.sc`, also in the `refund` package, which is among the available implementations. Next, we reestimate the mean $\mu_{ij}(t)$ in model (2.2) with function-on-scalar regression using `pffr`, although now, instead of assuming independence, we decompose the residuals using the principal components and score variances estimated in the previous step. We then reestimate principal components and scores using `fpca.sc`. The final step is to model the score variances given these score estimates. Assuming that the scores are normally distributed conditional on random effects and covariates, model (2.3) induces a generalized gamma linear mixed model for ξ_{ijk}^2, the square of the scores, with log link, coefficients γ_{lk} and g_{imk}, and shape parameter equal to $1/2$. We fit this model with the `lme4` package, separately with respect to the scores for each principal component, in order to obtain estimates of our parameters of interest in the score variance model [Bates *et al.*, 2015].

The first two steps of this approach are consistent with the common strategy for FPCA, and we account for non-constant score variance through an additional modeling step. We anticipate that this sequential approach will work reasonably well in many cases, but note that it arises as a sequence of models that treat estimated quantities as fixed. First, one estimates the mean; then one treats the mean as fixed to estimate the principal components and the scores; finally, one treats the scores as fixed to estimate the score variance model. Overall performance may deteriorate by failing to incorporate uncertainty in estimates in each step, particularly in cases of sparsely observed curves or high measurement error variances [Goldsmith *et al.*, 2013]. Additionally, because scores are typically estimated in a mixed model framework, the use of *marginal* score variances in the FPCA step can negatively impact score estimation and the subsequent modeling of *conditional* score variances.

2.4.2 Bayesian approach

2.4.2.1 Bayesian model

Jointly estimating all parameters in models (2.2) and (2.3) in a Bayesian framework is an appealing alternative to the sequential estimation approach. We expect this to be less familiar to readers than the sequential approach, and therefore provide a more detailed description.

Our Bayesian specification of these models is formulated in matrix form to reflect the discrete nature of the observed data. In the following $\boldsymbol{\Theta}$ is a known $D \times K_\theta$ matrix of K_θ spline basis functions evaluated on the shared grid of length D on which the curves are observed. We assume a normal distribution of the scores ξ_{ijk} conditional on random effects and covariates:

$$
\begin{aligned}
\boldsymbol{p}_{ij} &= \sum_{l=0}^{L} x_{ijl}\boldsymbol{\Theta}\boldsymbol{\beta}_l + \sum_{m=1}^{M} z_{ijm}\boldsymbol{\Theta}\boldsymbol{b}_{im} + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \qquad (2.5)\\
\boldsymbol{\beta}_l &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\beta}_l}\boldsymbol{Q}^{-1}\right]; \sigma^2_{\boldsymbol{\beta}_l} \sim \text{IG}\left[\alpha, \beta\right]\\
\boldsymbol{b}_i &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{b}}((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})^{-1}\right]; \sigma^2_{\boldsymbol{b}} \sim \text{IG}\left[\alpha, \beta\right]\\
\boldsymbol{\phi}_k &\sim \text{MVN}\left[0, \sigma^2_{\boldsymbol{\phi}_k}\boldsymbol{Q}^{-1}\right]; \sigma^2_{\boldsymbol{\phi}_k} \sim \text{IG}\left[\alpha, \beta\right]\\
\xi_{ijk} &\sim \text{N}\left[0, \exp\left(\sum_{l=0}^{L^*} \gamma_{lk} x^*_{ijlk} + \sum_{m=1}^{M^*} g_{imk} z^*_{ijmk}\right)\right]\\
\gamma_{lk} &\sim \text{N}\left[0, \sigma^2_{\gamma_{lk}}\right]\\
\boldsymbol{g}_{ik} &\sim \text{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_k}\right]; \boldsymbol{\Sigma}_{\boldsymbol{g}_k} \sim \text{IW}\left[\boldsymbol{\Psi}_k, \nu\right]\\
\boldsymbol{\epsilon}_{ij} &\sim \text{MVN}\left[0, \sigma^2\boldsymbol{I}\right]; \sigma^2 \sim \text{IG}\left[\alpha, \beta\right]
\end{aligned}
$$

In model (2.5), $i = 1, \ldots, I$ refers to subjects, $j = 1, \ldots, J_i$ refers to motions within subjects, and $k = 1, \ldots, K$ refers to principal components. We define the total number of functional observations $n = \sum_{i=1}^{I} J_i$. The column vectors $\boldsymbol{p}_{ij}$ and $\boldsymbol{\epsilon}_{ij}$ are the $D \times 1$ observed functional outcome and independent error term, respectively, on the finite grid shared across subjects for the jth curve of the ith subject. The vectors $\boldsymbol{\beta}_l$, for $l = 0, \ldots, L$, are functional effect spline coefficient vectors, $\boldsymbol{b}_{im}$, for $i = 1, \ldots, I$ and $m = 1, \ldots, M$, are random effect spline coefficient vectors, and $\boldsymbol{\phi}_k$, for $k = 1 \ldots, K$, are principal component spline coefficient vec-

tors, all of length K_θ. $\boldsymbol{Q}$ is a penalty matrix of the form $\boldsymbol{\Theta}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{\Theta}$, where $\boldsymbol{M}$ is a matrix that penalizes the second derivative of the estimated functions. $\boldsymbol{I}$ is the identity matrix. MVN refers to the multivariate normal distribution, N to the normal distribution, IG to the inverse-gamma distribution, and IW to the inverse-Wishart distribution. Models (2.3) and (2.4) can be written in the form of model (2.5) above by introducing into those models covariates x^*_{ij0k} (in model (2.3), multiplying γ_{0k}) and x_{ij0} (in model (2.4), multiplying $\beta_0(t)$), identically equal to 1. Some of the models used here, like in our real data analysis, do not have a global functional intercept $\boldsymbol{\beta}_0$ or global score variance intercepts γ_{0k}; in these models there are no such covariates identically equal to 1.

As discussed further in Section 2.4.2.3, for purposes of identifiability and to obtain FPCs that represent non-overlapping directions of variation, when fitting this model we introduce the additional constraint that the FPCs should be orthonormal and that each FPC should explain the largest possible amount of variance in the data, conditionally on the previously estimated FPCs, if any.

In keeping with standard practice, we set the prior variances $\sigma^2_{\gamma_{lk}}$ for the fixed-effect coefficients in the score variance model to a large constant, so that their prior is close to uniform. We set ν, the degrees of freedom parameter for the inverse-Wishart prior for the covariance matrices $\Sigma_{\boldsymbol{g}_k}$, to the dimension of $\boldsymbol{g}_{ik}$. We use an empirical Bayes approach, discussed further in Section 2.4.2.4, to specify $\boldsymbol{\Psi}_k$, the scale matrix parameters of these inverse-Wishart priors. When the random effects $\boldsymbol{g}_{ik}$ are one-dimensional, this prior reduces to an inverse-Gamma prior. Sensitivity to prior specifications of this model should be explored, and we do so with respect to our real data analysis in Appendix A.4.

Variance components $\{\sigma^2_{\boldsymbol{\beta}_l}\}_{l=0}^L$ and $\{\sigma^2_{\boldsymbol{\phi}_k}\}_{k=1}^K$ act as tuning parameters controlling the smoothness of coefficient functions $\beta_l(t)$ and FPC functions $\phi_k(t)$, and our prior specification for them is related to standard techniques in semiparametric regression. $\sigma^2_{\boldsymbol{b}}$, meanwhile, is a tuning parameter that controls the amount of penalization of the random effects, and is shared across the $b_{im}(t)$, so that all random effects for all subjects share a common distribution. Whereas fixed effects and functional principal components are penalized only through their squared second derivative, the magnitude of the random effects is also penalized through the full-rank penalty matrix $\boldsymbol{I}$ to ensure identifiability [Scheipl *et al.*, 2015; **?**].

The parameter π, $0 < \pi < 1$, determines the balance of smoothness and shrinkage penalties in the estimation of the random effects $b_{im}(t)$. We discuss how to set the value of this parameter in Section 2.4.2.4. We set α and β, the parameters of the inverse-gamma prior distributions for the variance components, to 1.

Our framework can accommodate more complicated random effect structures. In our application in Section 2.6, for example, each subject has 8 random effect vectors $\boldsymbol{g}_{ilk}$, one for each target, indexed by $l = 1, \ldots, 8$; the index l is used here since in Section 2.6 l is used to index targets. We model the correlations between these random effect vectors through a nested random effect structure:

$$\boldsymbol{g}_{ilk} \sim \text{MVN}\left[\boldsymbol{g}_{ik}, \boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}\right]; \qquad \boldsymbol{g}_{ik} \sim \text{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_{k}}\right] \tag{2.6}$$

Here the random effect vectors $\boldsymbol{g}_{ilk}$ for subject i and FPC k, $l = 1, \ldots 8$, are centered around a subject-specific random effect vector $\boldsymbol{g}_{ik}$. We estimate two separate random effect covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_{k}}$, for each FPC k, one at the subject-target level and one at the subject level. These matrices are given inverse-Wishart priors, and are discussed further in Section 2.4.2.4.

2.4.2.2 Estimation strategies

Sampling-based approaches to Bayesian inference of model (2.5) are challenging due to the constraints we impose on the $\phi_k(t)$ for purposes of interpretability of the score variance models, which are our primary interest. We present two methods for Bayesian estimation and inference for model (2.5): first, an iterative variational Bayes method, and second, a Hamiltonian Monte Carlo (HMC) sampler, implemented with the `STAN` Bayesian programming language [Stan Development Team, 2013]. Our iterative variational Bayes method, which estimates each parameter in turn conditional on currently estimated values of the other parameters, is described in detail in Appendix A.5. This appendix also includes a brief overview of variational Bayes methods. Our HMC sampler, also described in Appendix A.5, conditions on estimates of the FPCs and fixed and random functional effects from the variational Bayes method, and estimates the other quantities in model (2.5).

2.4.2.3 Orthonormalization

A well-known challenge for Bayesian and probabilistic approaches to FPCA is that the basis functions $\phi_k(t)$ are not constrained to be orthogonal. In addition, when the scores ξ_{ijk} do not have unit variance, the basis functions will also be indeterminate up to magnitude, since any increase in their norm can be accommodated by decreased variance of the scores. Where interest lies in the variance of scores with respect to particular basis functions, it is important for the basis functions to be well-identified and orthogonal, so that they represent distinct and non-overlapping modes of variation. We therefore constrain estimated FPCs to be orthonormal and require each FPC to explain the largest possible amount of variance in the data, conditionally on the previously estimated FPCs, if any.

Let $\boldsymbol{\Xi}$ be the $n \times K$ matrix of principal component scores and $\boldsymbol{\Phi}$ the K by K_θ matrix of principal component spline coefficient vectors. In each step of our iterative variational Bayes algorithm, we apply the singular value decomposition to the matrix product $\boldsymbol{\Xi}\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T$; the orthonormalized principal component basis vectors which satisfy these constraints are then the right singular vectors of this decomposition. A similar approach was used to induce orthogonality of the principal components in the Monte Carlo Expectation Maximization algorithm of [Huang *et al.*, 2014] and as a post-processing step in [Goldsmith *et al.*, 2015]. Although explicit orthonormality constraints may be possible in this setting [Šmídl and Quinn, 2007], our simple approach, while not exact, provides for accurate estimation. Our HMC sampler conditions on the variational Bayes estimates of the FPCs, and therefore also satisfies the desired constraints.

2.4.2.4 Hyperparameter selection

The parameter π in model (2.5) controls the balance of smoothness and shrinkage penalization in the estimation of the random effects $\boldsymbol{b}_{im}$. In our variational Bayes approach we choose π to minimize the Bayesian information criterion [**?**], following the approach of **?**.

To set the hyperparameter $\boldsymbol{\Psi}_k$ in model (2.5) (or the hyperparameters in the inverse-Wishart priors for the variance parameters in model (2.6)), we use an empirical Bayes method. First, we estimate scores ξ_{ijk} using our variational Bayes method, with a constant score variance for each FPC. We then estimate the random effects $\boldsymbol{g}_{ik}$ (or $\boldsymbol{g}_{ilk}$) using a

generalized gamma linear mixed model, as described in Section 2.4.1. Finally, we compute the empirical covariance matrix corresponding to $\mathbf{\Sigma}_{\boldsymbol{g}_k}$ (or $\mathbf{\Sigma}_{\boldsymbol{g}_{ik}}$ and $\mathbf{\Sigma}_{\boldsymbol{g}_k}$), and set the hyperparameter so that the mode of the prior distribution matches this empirical covariance matrix.

2.5 Simulations

We demonstrate the performance of our method using simulated data. Here we present a simulation that includes functional random effects as well as scalar score variance random effects. Appendix A.6 includes additional simulations in a cross-sectional context which demonstrate the effect of varying the number of estimated FPCs, the number of spline basis functions, and the measurement error.

In our simulation design, the jth curve for the ith subject is generated from the model

$$P_{ij}(t) = 0 + b_i(t) + \sum_{k=1}^{4} \xi_{ijk}\phi_k(t) + \epsilon_{ij}(t) \tag{2.7}$$

We observe the curves at $D = 50$ equally spaced points on the domain $[0, 2\pi]$. FPCs ϕ_1 and ϕ_2 correspond to the functions $\sin(x)$ and $\cos(x)$ and FPCs ϕ_3 and ϕ_4 correspond to the functions $\sin(2x)$ and $\cos(2x)$. We divide the curves equally into two groups $m = 1, 2$. We define x^*_{ij1} to be equal to 1 if the ith subject is assigned to group 1, and 0 otherwise, and we define x^*_{ij2} to be equal to 1 if the ith subject is assigned to group 2, and 0 otherwise. We generate scores ξ_{ijk} from zero-mean normal distributions with variances equal to

$$\text{Var}(\xi_{ijk}|\boldsymbol{x}^*_{ij}, g_{ik}) = \exp\left(\sum_{l=1}^{2} \gamma_{lk} x^*_{ijl} + g_{ik}\right) \tag{2.8}$$

We set γ_{1k} for $k = 1, \ldots, 4$ to the natural logarithms of $36, 12, 6$ and 4, respectively, and γ_{2k} for $k = 1, \ldots, 4$ to the natural logarithms of $18, 24, 12$ and 6, respectively. The order of γ_{1k} and γ_{2k} for FPCs (represented by k) 1 and 2 black is purposely reversed between groups 1 and 2 so that the dominant mode of variation is not the same in the two groups. We generate the random effects g_{ik} in the score variance model from a normal distribution with mean zero and variance $\sigma^2_{g_k}$, setting $\sigma^2_{g_k}$ to 3.0, 1.0, 0.3, and 0.1 across FPCs. We simulate functional random effects $b_i(t)$ for each subject by generating 10 elements of a random

effect spline coefficient vector from the distribution MVN $\left[0, \sigma^2_{\boldsymbol{b}}((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})^{-1}\right]$, and then multiplying this vector by a B-spline basis function evaluation matrix. We set $\pi = \sigma^2_{\boldsymbol{b}} = 1/2000$, resulting in smooth random effects approximately one-third the magnitude of the FPC deviations. The $\epsilon_{ij}(t)$ are independent errors generated at all t from a normal distribution with mean zero and variance $\sigma^2 = 0.25$.

We fix the sample size I at 24 and set the number of curves per subject J_i to 4, 12, 24 and 48. Two hundred replicate datasets were generated for each of the four scenarios. The simulation scenario with $I = J_i = 24$ is closest to the sample size in our real data application, where for each of 8 targets we have $I = 26$ and $J_i \approx 24$.

We fit the following model to each simulated dataset using each of the three approaches described in Section 3.2:

$$\begin{aligned} \boldsymbol{p}_{ij} &= \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \boldsymbol{\Theta}\boldsymbol{b}_i + \sum_{k=1}^{4} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \\ \xi_{ijk} &\sim \mathrm{N}\left[0, \exp\left(\sum_{l=1}^{2} \gamma_{lk} x^*_{ijl} + g_{ik}\right)\right]. \end{aligned}$$

Here $\boldsymbol{p}_{ij}$ is the vectorized observation of $P_{ij}(t)$ from model (2.7). We use 10 spline basis functions for estimation, so that $\boldsymbol{\Theta}$ is a 50×10 B-spline basis function evaluation matrix. For the Bayesian approaches, we use the priors specified in model (2.5), including $\mathrm{N}[0, 100]$ priors for variance parameters $\sigma^2_{\gamma_{lk}}$. We use the empirical Bayes approach discussed in Section 2.4.2.4 to set the scale parameters for the inverse-gamma priors for the variances $\sigma^2_{g_k}$ of the random effects g_{ik}.

Figures 2.3, 2.4 and 2.5 illustrate the quality of variational Bayes (VB) estimation of functional random effects, FPCs, and fixed and random effect score variance parameters. The top row of Figure 2.3 shows the collection of simulated curves for two subjects and includes the true and estimated subject-specific mean. The bottom row of this figure shows the true and estimated score variances across FPCs for a single simulated dataset, and suggests that fixed and random effects in the score variance model can be well-estimated.

The top row of Figure 2.4 shows estimated FPCs across all simulated datasets with $J_i = 24$; the FPCs are well-estimated and have no obvious systematic biases. The bottom row shows integrated squared errors (ISEs) for the FPCs across each possible J_i. As expected,

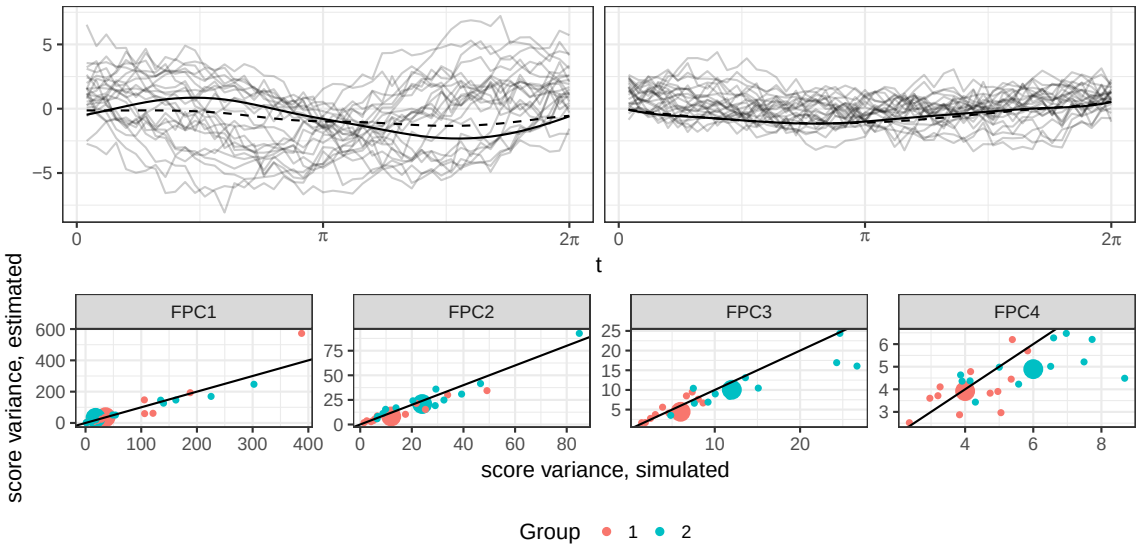

Figure 2.3: Selected results for the VB method for one simulation replicate with $I = J_i = 24$. This simulation replicate was selected because the estimation quality of the group-level score variances, shown in the bottom row, is close to median with respect to all simulations. Panels in the top row show simulated curves for two subjects in light black, the simulated functional random effect for that subject as a dashed line, and the estimated functional random effect for that subject as a dark solid line. The subjects were selected to show one subject with a poorly estimated functional random effect (left) and one with a well estimated functional random effect (right). Panels in the bottom row show, for each FPC, estimates and simulated values of the group-level and subject-specific score variances. Large colored dots are the group-level score variances, and small colored dots are the estimated score variances for each subject, i.e., they combine the fixed effect and the random effect.

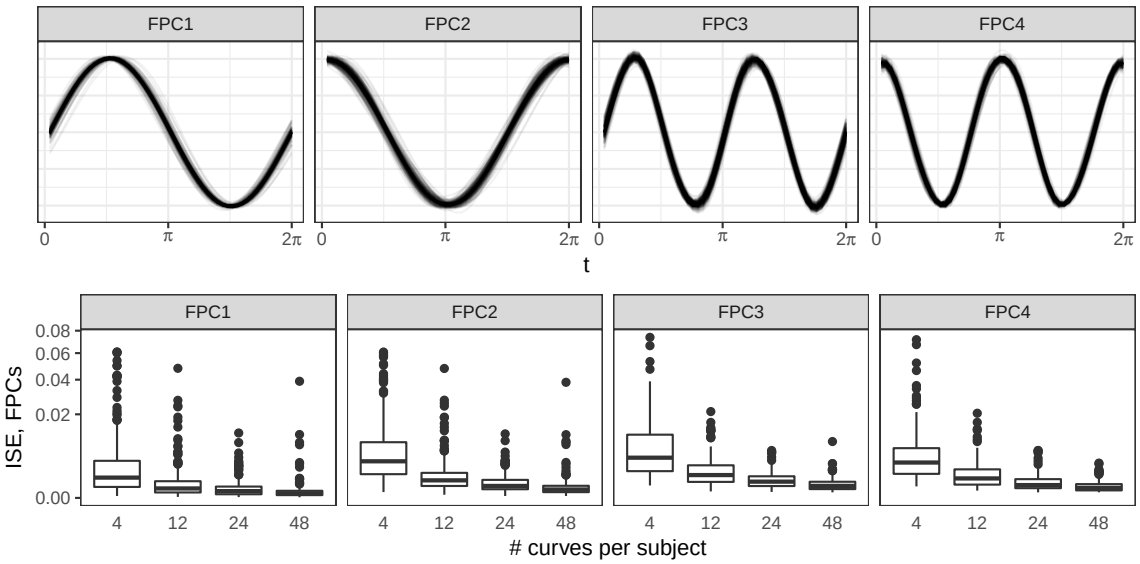

Figure 2.4: Estimation of FPCs using the VB method. Panels in the top row show a true FPC in dark black, and the VB estimates of that FPC for all simulation replicates with $J_i = 24$ in light black. Panels in the bottom row show, for each FPC and J_i, boxplots of integrated square errors (ISEs) for VB estimates $\widehat{\phi_k}(t)$ of each FPC $\phi_k(t)$, defined as $\text{ISE} = \int_0^{2\pi}[\phi_k(t) - \widehat{\phi_k}(t)]^2 dt$. The estimates in the top row therefore correspond to the ISEs for $J_i = 24$ shown in the bottom row. Figure A.10 in Appendix A.6 shows examples of estimates of FPCs with a range of different ISEs.

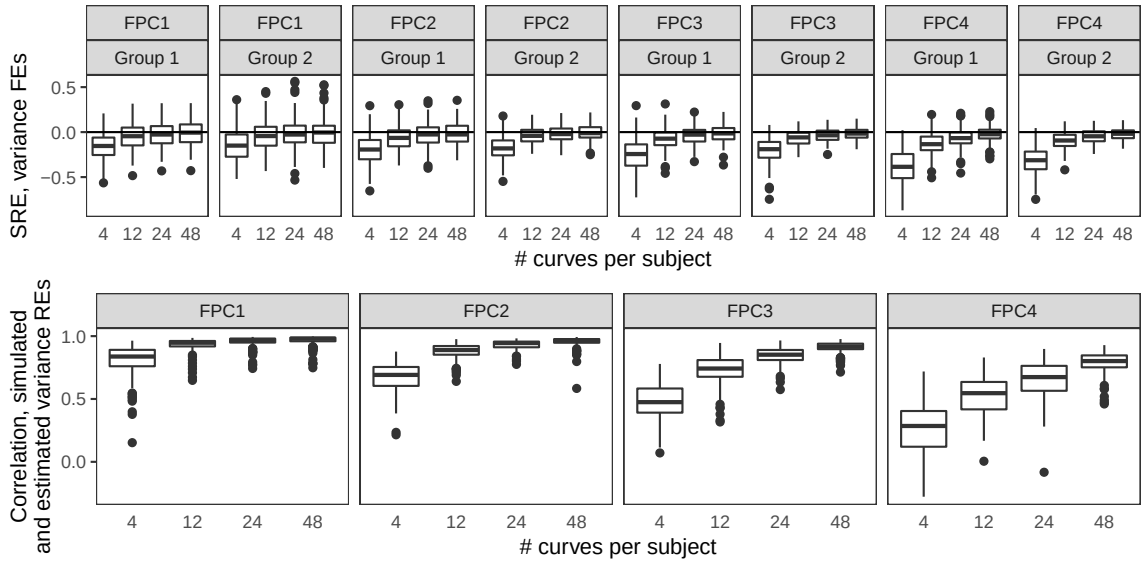

Figure 2.5: Estimation of score variance fixed and random effects using VB. Panels in the top row show, for each FPC, group, and J_i, boxplots of signed relative errors (SREs) for VB estimates $\widehat{\gamma_{lk}}$ of the fixed effect score variance parameters γ_{lk}, defined as $\text{SRE} = \frac{\widehat{\gamma_{lk}} - \gamma_{lk}}{\gamma_{lk}}$. Panels in the bottom row show, for each FPC and J_i, the correlation between random effect score variance parameters g_{ik} and their VB estimates. Intercepts and slopes for linear regressions of estimated on simulated random effect score variances are centered around 0 and 1, respectively (not shown).

the ISEs are smaller for the FPCs with larger score variances, and decrease as J_i increases. For 12 and especially for 4 curves per subject, estimates of the FPCs correspond to linear combinations of the simulated FPCs, leading to high ISEs and to inaccurate estimates of parameters in our score variance model (examples of poorly estimated FPCs can be seen in Appendix A.6).

Panels in the top row of Figure 2.5 show that estimates of fixed effect score variance parameters are shrunk towards zero, especially for lower numbers of curves per subject and FPCs 3 and 4. We attribute this to overfitting of the random effects in the mean model, which incorporates some of the variability attributable to the FPCs into the estimated random effects and reduces estimated score variances. Score variance random effects, shown in the bottom row of Figure 2.5, are more accurately estimated with more curves per subject.

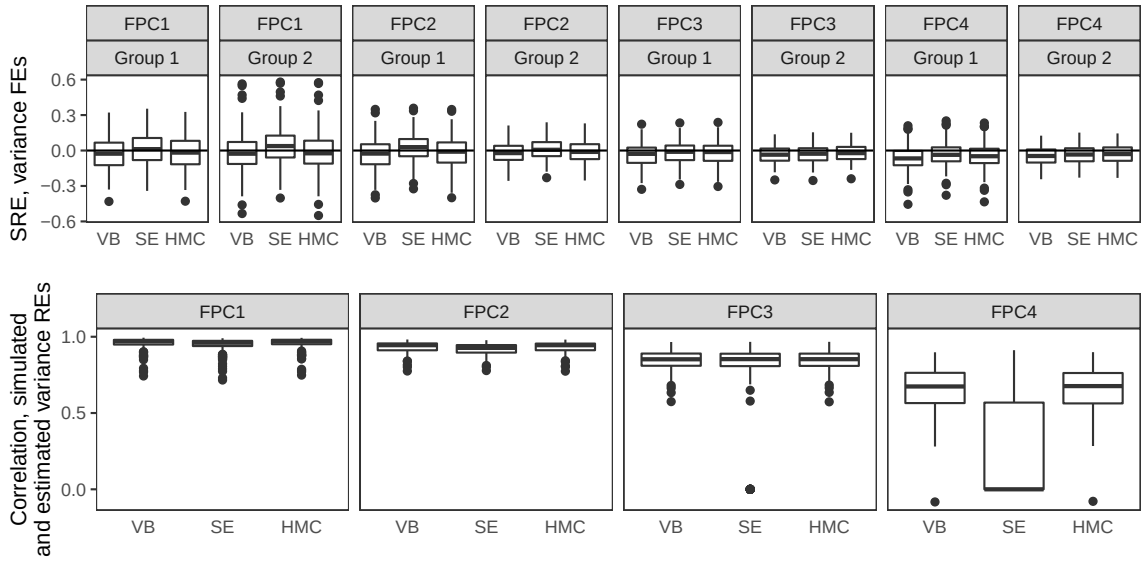

Figure 2.6: Comparison of estimation of score variance fixed and random effects using three methods. Panels in the top row show, for each FPC, group, and estimation method, boxplots of signed relative errors (SREs) for estimates of the fixed effect score variance parameters γ_{lk} for $J_i = 24$. Panels in the bottom row show, for each FPC and estimation method, the correlation between random effect score variance parameters g_{ik} and their estimates for $J_i = 24$. Intercepts and slopes for linear regressions of estimated on simulated random effect score variances are centered around 0 and 1, respectively (not shown).

Figure 2.6 and Table 2.1 show results from a comparison of the VB estimation procedure to the sequential estimation (SE) and Hamiltonian Monte Carlo (HMC) methods described in Section 3.2. We ran 4 HMC chains for 800 iterations each, and discarded the first 400 iterations from each chain. We assessed convergence of the chains by examining the convergence criterion of Gelman and Rubin [1992]. Values of this criterion near 1 indicate convergence. For each of our simulation runs the criterion for every sampled variable was less than 1.1, and usually much closer to 1, suggesting convergence of the chains. In general, performance for the VB and HMC methods is comparable, and both methods are in some respects superior to the performance of the SE method. Figure 2.6 compares the three methods' estimation of the score variance parameters. Especially for FPC 4, the SE method occasionally estimates random effect variances at 0; these are represented in the lower-right panel of Figure 2.6 as points where the correlation between simulated and estimated score variance random effects is 0. Table 2.1 shows, based on the simulation scenario with $J_i = 24$, the frequentist coverage of 95% credible intervals for the VB and HMC methods, and of 95% confidence intervals for the SE method, in each case, for the fixed effect score variance parameters γ_{lk}. For FPCs 3 and 4 especially, the SE procedure confidence intervals are too narrow. The median ISE for the functional random effects is about 30% higher with the VB method than with the SE method. This results from the relative tendency of the VB method to shrink FPC score estimates to zero; when the mean of the scores is in fact non-zero, this shifts estimated functional random effects away from zero. Other comparisons of these methods are broadly similar.

The HMC method is more computationally expensive than the other two methods. Running 4 chains for 800 iterations in parallel took approximately 90 minutes for $J_i = 24$. On one processor, by comparison, the SE method took about 20 minutes, almost entirely to run function-on-scalar regression using `pffr`. The VB method took approximately six minutes, including the grid search to set the value of the parameter π, which controls the balance between zeroth and second-derivative penalties in the estimation of functional random effects.

FPC	Group	VB	SE	HMC
1	1	0.955	0.915	0.960
1	2	0.945	0.905	0.945
2	1	0.940	0.935	0.940
2	2	0.980	0.935	0.975
3	1	0.965	0.930	0.975
3	2	0.955	0.885	0.980
4	1	0.930	0.775	0.970
4	2	0.940	0.705	0.965

Table 2.1: Coverage of 95% credible/confidence intervals for the score variance parameters γ_{lk} using the VB, SE and HMC procedures, for $J_i = 24$.

2.6 Analysis of kinematic data

We now apply the methods described above to our motivating dataset. To reiterate, our goal is to quantify the process of motor learning in healthy subjects, with a focus on the reduction of motor variance through repetition. Our dataset consists of 26 healthy, right-handed subjects making repeated motions to each of 8 targets. We focus on estimation, interpretation and inference for the parameters in a covariate and subject-dependent heteroskedastic FPCA model, with primary interest in the effect of repetition number in the model for score variance. We hypothesize that variance will be lower for later repetitions due to skill learning.

Prior to fitting the model, we rotate all motions to be in the direction of the target at 0° so that the X axis is the major axis of motion. For this reason, variation along the X axis is interpretable as variation in motion extent and variation along the Y axis is interpretable as variation in motion direction. We present results for univariate analyses of the $P_{ij}^X(t)$ and $P_{ij}^Y(t)$ position curves in the right hand and describe a bivariate approach to modeling the same data.

We present models with 2 FPCs, since 2 FPCs are sufficient to explain roughly 95% of the motion variability (and usually more) of motions remaining after accounting for fixed

and random effects in the mean structure. Most of the variability of motions around the mean is explained by the first FPC, so we emphasize score variance of the first FPC as a convenient summary for the motion variance, and briefly present some results for the second FPC.

2.6.1 Model

We examine the effect of practice on the variance of motions while accounting for target and individual-specific idiosyncrasies. To do this, we use a model for score variance that includes a fixed intercept and slope parameter for each target and one random intercept and slope parameter for each subject-target combination. Correlation between score variance random effects for different targets for the same subject is induced via a nested random effects structure. The mean structure for observed curves consists of functional intercepts $\boldsymbol{\beta}_l$ for each target $l \in \{1, \ldots, 8\}$ and random effects $\boldsymbol{b}_{il}$ for each subject-target combination, to account for heterogeneity in the average motion across subjects and targets. Our heteroskedastic FPCA model is therefore:

$$\boldsymbol{p}_{ij} = \sum_{l=1}^{8} \mathbb{I}(tar_{ij} = l)\left(\boldsymbol{\Theta}\boldsymbol{\beta}_l + \boldsymbol{\Theta}\boldsymbol{b}_{il}\right) + \sum_{k=1}^{K} \xi_{ijk}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_{ij} \tag{2.9}$$

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{l=1}^{8} \mathbb{I}(tar_{ij} = l)\left(\gamma_{lk,int} + g_{ilk,int} + (rep_{ij} - 1)(\gamma_{lk,slope} + g_{ilk,slope})\right)\right)\right] \tag{2.10}$$

$$\boldsymbol{g}_{ilk} \sim \mathrm{MVN}\left[\boldsymbol{g}_{ik}, \boldsymbol{\Sigma}_{\boldsymbol{g}_{ik}}\right];\ \boldsymbol{g}_{ik} \sim \mathrm{MVN}\left[0, \boldsymbol{\Sigma}_{\boldsymbol{g}_k}\right]$$

The covariate tar_{ij} indicates the target to which motion j by subject i is directed. The covariate rep_{ij} indicates the repetition number of motion j, starting at 1, among all motions by subject i to the target to which motion j is directed, and $\mathbb{I}(\cdot)$ is the indicator function. To accommodate differences in baseline variance across targets, this model includes separate population-level intercepts $\gamma_{lk,int}$ for each target l. The slopes $\gamma_{lk,slope}$ on repetition number indicate the change in variance due to practice for target l; negative values indicate a reduction in motion variance. To accommodate subject and target-specific effects, each subject-target combination has a random intercept $g_{ilk,int}$ and a random slope $g_{ilk,slope}$, and each subject has an overall random intercept $g_{ik,int}$ and overall random slope $g_{ik,slope}$, in the

score variance model for each functional principal component. This model parameterization allows different baseline variances and changes in variance for each target and subject, but shares FPC basis functions across targets. The model also assumes independence of functional random effects $\boldsymbol{b}_{il}$, $l = 1, \ldots, 8$ by the same subject to different targets, as well as independence of functional random effects $\boldsymbol{b}_{il}$ and score variance random effects $\boldsymbol{g}_{ilk}$ for the same subject. The validity of these assumptions for our data are discussed in Appendix A.4.

Throughout, fixed effects $\gamma_{lk,int}$ and $\gamma_{lk,slope}$ are given $\mathrm{N}[0, 100]$ priors. Random effects $g_{ilk,int}$ and $g_{ilk,slope}$ are modeled using a bivariate normal distribution to allow for correlation between the random intercept and slope parameters in each FPC score variance model, and with nesting to allow for correlations between the random effects for the same subject and different targets. We use the empirical Bayes method described in Section 2.4.2.4 to set the scale matrix parameters of the inverse-Wishart priors for $\boldsymbol{g}_{ilk}$ and $\boldsymbol{g}_{ik}$. Appendix A.4 includes an analysis which examines the sensitivity of our results to various choices of prior hyperparameters.

We fit (2.9) and (2.10) using our VB method, with $K = 2$ principal components and a cubic B-spline evaluation matrix $\boldsymbol{\Theta}$ with $K_\theta = 10$ basis functions.

2.6.2 Results

Figure 2.7 shows estimated score variances as a function of repetition number for X and Y coordinate right hand motions to all targets. There is a decreasing trend in score variance for the first principal component scores for all targets and for both the X and Y coordinates, which agrees with our hypotheses regarding learning. Figure 2.7 also shows that nearly all of the variance of motion is attributable to the first FPC. Baseline variance is generally higher in the X direction than the Y direction, indicating that motion extent is generally more variable than motion direction.

To examine the adequacy of modeling score variance as a function of repetition number with a linear model, we compared the results of model (2.10) with a model for the score variances saturated in repetition number, i.e., where each repetition number m has its own

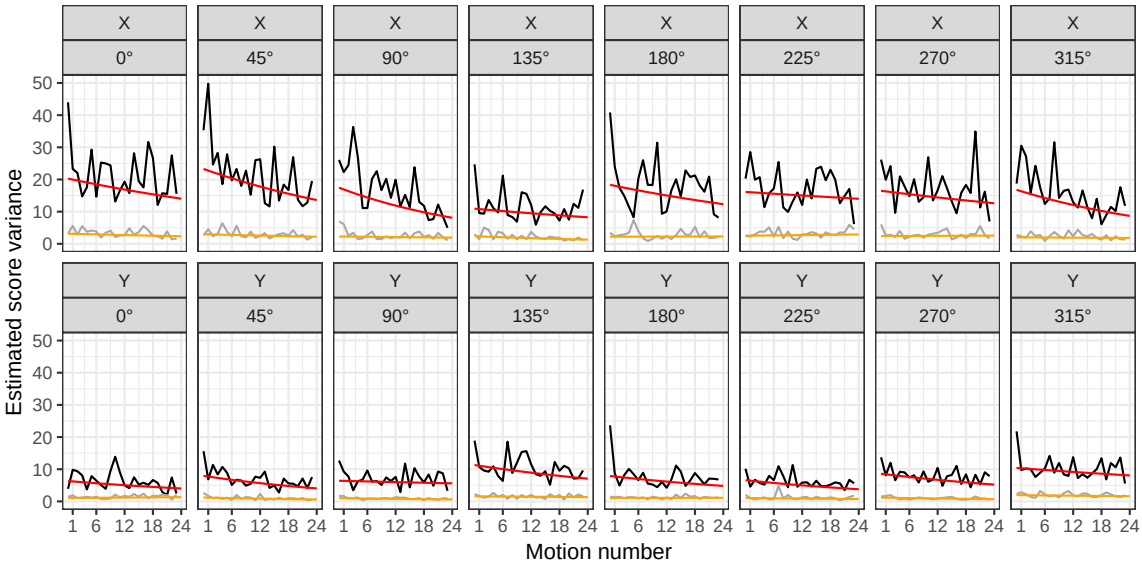

Figure 2.7: VB estimates of score variances for right hand motions to each target (in columns), separately for each direction (X or Y, in rows). Panels show the VB estimates of the score variance as a function of repetition number using the slope-intercept model (2.10) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (2.11), in black and grey (first and second FPC, respectively).

set of parameters γ_{lkm} in the model for the score variances:

$$\xi_{ijk} \sim \mathrm{N}\left[0, \sigma^2_{\xi_{ijk}} = \exp\left(\sum_{l=1}^{8}\sum_{m=1}^{24} \mathbb{I}(tar_{ij} = l, rep_{ij} = m)\gamma_{lkm}\right)\right]. \tag{2.11}$$

The results for these two models are included in Figure 2.7. The general agreement between the linear and saturated models suggests that the slope-intercept model is reasonable. For some targets score variance is especially high for the first motion, which may reflect a familiarization with the experimental apparatus.

We now consider inference for the decreasing trend in variance for the first principal component scores. We are interested in the parameters $\gamma_{l1,slope}$, which estimate the population-level target-specific changes in score variance for the first principal component with each additional motion. Figure 2.8 shows VB estimates and 95% credible intervals for the $\gamma_{l1,slope}$ parameters for motions by the right hand to each target. All the point estimates

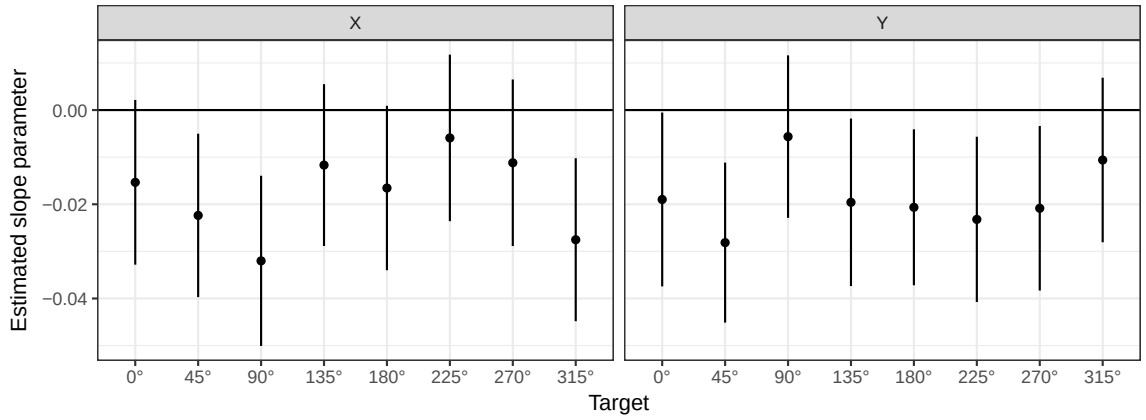

Figure 2.8: VB estimates of $\gamma_{l1,slope}$. This figure shows VB estimates and 95% credible intervals for target-specific score variance slope parameters $\gamma_{l1,slope}$ for motions by the right hand to each target, for the X and Y coordinates.

$\gamma_{l1,slope}$ are lower than 0, indicating decreasing first principal component score variance with additional repetition. For some targets and coordinates there is substantial evidence that $\gamma_{l1,slope} < 0$; these results are consistent with our understanding of motor learning, although they do not adjust for multiple comparisons.

Appendix A.3 includes results of a bivariate approach to modeling motion kinematics, which accounts for the 2-dimensional nature of the motions. In this approach, the X and Y coordinates of curves are concatenated, and each principal component reflects variation in both X and Y coordinates. For curves rotated to extend in the same direction, the results of this approach suggests that variation in motion extent (represented by the X coordinate) and motion direction (represented by the Y coordinate) are largely uncorrelated: the estimate of the first bivariate FPC represents variation primarily in the X coordinate, and is similar to the estimate of the first FPC in the X coordinate model, and vice versa for the second bivariate FPC. Analyses of score variance, then, closely follow the preceding univariate analyses.

Appendix A.2 includes an analysis of data for one target using the VB, HMC and SE methods. The three methods yield similar results.

2.7 Discussion

This chapter develops a framework for the analysis of covariate and subject-dependent patterns of motion variance in kinematic data. Our methods allow for flexible modeling of the covariate-dependence of variance of functional data with easily interpretable results. Our approach allows for the estimation of subject-specific effects on variance, as well as the consideration of multiple covariates.

By applying these methods to our motivating dataset, we have demonstrated that motion variance is reduced with repetition. Results in Appendix A.1 additionally show that the baseline level of skill of subjects is correlated across targets and hands, and that baseline variance is considerably greater in the non-dominant than the dominant hand. Further applications of these methods in scientifically important contexts could focus, for example, on whether motion variance is reduced with training faster in the dominant hand, or on whether training with one hand transfers skill to the other hand. Further research could also investigate target-specific differences in improvement of variance with training. Movements to some of the targets require coordination between the shoulder and elbow, whereas others are primarily single-joint motions; the effectiveness of training may depend on the complexity of the motion.

We have provided three different estimation approaches for fitting heteroskedastic functional principal components models. Given its computational efficiency and comparable accuracy to the HMC and SE methods, we recommend use of the VB approach for exploratory analyses and model building. However, because of its approximate nature, we advise that any conclusions derived from the VB approach be confirmed with one of the other two methods, perhaps with a subset of the data if required for computational feasibility.

An alternative approach to the analysis of this dataset could treat the target effects $\gamma_{lk,int}$ and $\gamma_{lk,slope}$ in model (2.10) for the score variances as random effects centered around parameters $\mu_{k,int}$ and $\mu_{k,slope}$, representing the average across-target baseline score variance and change in score variance with repetition. Some advantages of this approach would be the estimation of parameters that summarize the global effect of repetition on motion variance and shrinkage of the target-specific score variance parameters. However, with

only 8 random target effects, the model would be sensitive to the specification of priors. Moreover, as discussed above, motions to different targets impose different demands on coordination and skill, which may reduce the interpretability of the parameters $\mu_{k,int}$ and $\mu_{k,slope}$.

Our analysis here is of curves linearly registered onto a common time domain, although our method could be applied to curves with different time domains, or to sparsely observed functional data. Current ongoing research will yield an improved approach to registration in kinematic experiments which will take account of the repeated observations at the subject level by seeking to estimate subject- and curve-specific warping functions. This approach, combined with the methods we present in this chapter, will eventually allow a more complete model for motion variability that takes into account both variability in motion duration and variability in motion trajectories.

There are several directions for further development. A full Bayesian treatment could estimate all quantities in model (2.5) jointly, or could condition on only the FPCs and jointly estimate all other quantities; given the very flexible nature of this model, additional constraints might be required in such a Bayesian treatment to improve identifiability. More complex models could allow for correlations between functional random effects and score variance random effects. Considering our data from the perspective of shape analysis may provide better understanding of interpretable motion features like location, scale and orientation [Kurtek *et al.*, 2012; Gu *et al.*, 2012]. Lastly, an alternative approach to that presented here would be to model covariate-dependent score distributions through quantile regression. This may produce valuable insights into the complete distribution of motions, especially when this is not symmetric, but some work is needed to understand the connection of this technique to traditional FPCA.

Chapter 3

Non-negative matrix factorization approach to analysis of functional data

3.1 Scientific Motivation and Statistical Background

Accelerometers can be used to study human activity in an unbiased and continuous manner at high temporal resolution. As part of the Baltimore Longitudinal Study of Aging (BLSA) [Schrack *et al.*, 2014], a sample of elderly healthy subjects wore the Actiheart, a combined heart rate and accelerometer adhesively placed on the chest [Brage *et al.*, 2006]. The Actiheart measures physical activity every minute in activity counts, a cumulative summary of acceleration. There are multiple days of physical activity records for most subjects, providing a valuable resource to study patterns of activity in this sample of elderly Americans.

This BLSA dataset was previously analyzed in Goldsmith *et al.* [2015], using a generalized function-on-scalar regression model. That work analyzed the dataset in two different ways. In one analysis the count data were binarized, to represent either activity or inactivity, and continuous functional principal components (FPCs) and fixed effects on a latent scale, linked to the binary outcomes via the logit link function, were estimated. In an alter-

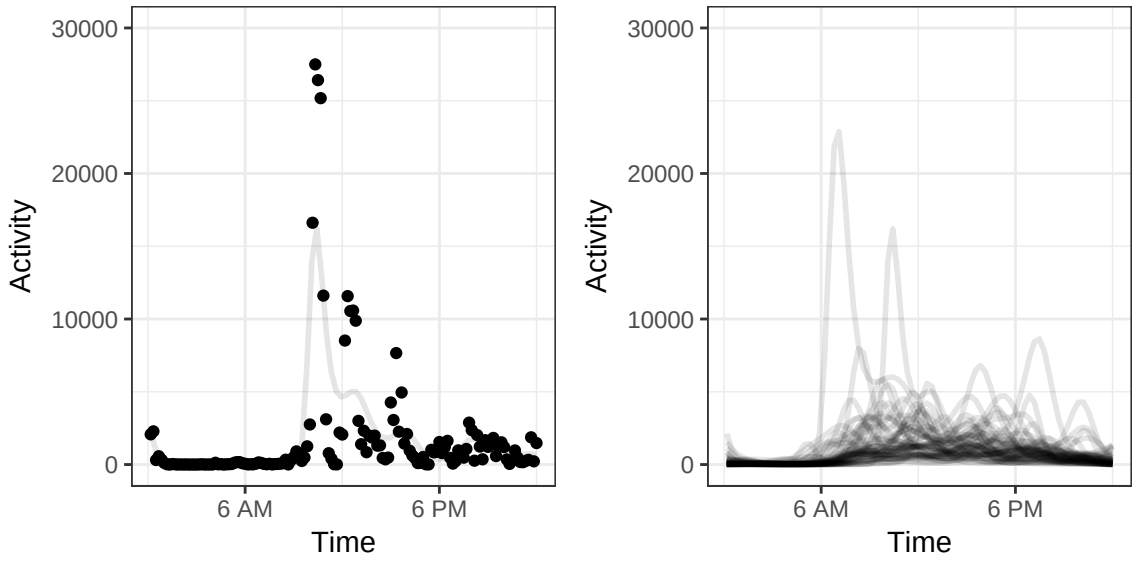

Figure 3.1: On the left is the raw data for one subject, showing activity summed over 5 days, binned in 10 minute intervals. A smooth of the data, fit using a generalized additive model with Poisson responses and a logarithmic link function with 15 basis functions, is also included. On the right are smooths for 50 subjects, including the subject shown on the left.

native analysis, the count data were treated as Poisson distributed, and continuous FPCs and fixed effects were estimated, again on a latent scale, linked to the observations via the logarithmic link function.

Here we use this dataset to present a novel decomposition of functional data, which results in decompositions that are more easily interpretable than those that result from modeling the data using exponential family distributions with canonical link functions. This decomposition draws on ideas from non-negative matrix factorization, and constrains both prototypic modes of variations and the coefficients encoding the decompositions of curves using these modes of variation to be non-negative. In non-negative matrix factorization [Lee and Seung, 1999], an $n \times m$ data matrix $\boldsymbol{Y}$, each column of which contains one observation, is approximated with a matrix product $\boldsymbol{V} \times \boldsymbol{H}$, where $\boldsymbol{V}$ and $\boldsymbol{H}$ are both non-negative rank r matrices. The $n \times r$ matrix $\boldsymbol{V}$ contains r different prototypes (these are also referred to in the literature as features, or basis images, among other terms), one in each column. Each column of the $r \times m$ matrix $\boldsymbol{H}$ encodes the contribution of each of the prototypes to the corresponding observation in $\boldsymbol{Y}$. The matrices $\boldsymbol{V}$ and $\boldsymbol{H}$ are often estimated via a scheme of multiplicative updates, for if the initial estimate of one of the coefficients is positive, and the update factors are also always positive, then the non-negativity constraint on the coefficients of the decomposition will be respected. Non-negative matrix factorization has been applied in a myriad of contexts, including in biostatistics [Sotiras *et al.*, 2015]. Conditions exist such that non-negative matrix factorization is unique, in that the data $\boldsymbol{Y}$ are only representable using one set of non-negative prototypes [Donoho and Stodden, 2003]. As shown in Section 3.4, in our simulations our method is able to correctly recover the functional prototypes used to generate simulated data.

In this chapter we extend non-negative matrix factorization to functional data analysis by expressing the prototypes in the columns of V in terms of spline basis functions, $V = \boldsymbol{\Theta} \times \boldsymbol{\Phi}$, where $\boldsymbol{\Theta}$ is a spline basis evaluation matrix. We encourage smoothness of the prototypes by penalizing their wiggliness with a second derivative penalty. In Section 3.2 below, we present this model and develop an estimation approach for it.

One benefit of our approach, in which functional prototypes are estimated on the data scale, without a link function transformation, is that the functional prototypes we estimate

are sparse and represent 'parts' that are transparently assembled into observations via addition, thus facilitating exploration of patterns of variation across subjects. This constrasts with approaches, like those of Hall *et al.* [2008], van der Linde [2009] and Goldsmith *et al.* [2015], that assume that observed data is related to a latent Gaussian process via a generalized linear model. In these models FPCs are interpretable and orthogonal on the latent logarithmic scale, not on the scale on which the data is observed. These approaches result in decompositions of observations that reflect highly complex patterns of cancellation and multiplication of non-sparse FPCs that often vary across their entire domain.

The remainder of this chapter is organized as follows. Section 3.2 presents our NARFD (Non-negative and Regularized Function Decomposition) model. Section 3.3 presents our implementation of a generalized FPC model for count data, to which we compare NARFD. Section 3.4 presents a simulation study that explores the estimation accuracy of NARFD. Section 3.5 presents our analysis of the BLSA accelerometer dataset. We close with a discussion.

3.2 Methods

We observe nonnegative integer data $Y_i(t_{ij})$ for subjects $i \in 1 \dots I$ and times $t_{11}, \dots, t_{IJ_i}$. Each subject is observed at a possibly unique set of J_i times $t_{i1}, t_{i2}, \dots, t_{iJ_i}$ in the domain $[T_1, T_2]$. In NARFD for Poisson data, we assume the following generative process for the trajectories:

$$\mu_i(t) = \sum_{k=1}^{K} \xi_{ik}\phi(t) \tag{3.1}$$

$$Y_i(t) \sim \text{Pois}\{\mu_i(t)\}$$

Here $\mu_i(t)$ is a latent trajectory, $\phi_k(t), k \in 1, \dots, K$, are functional prototypes, and ξ_{ik} are scores. Since $Y_i(t)$ has the Poisson distribution, the variance and mean of $Y_i(t)$ are both equal to $\mu_i(t)$. Both the scores ξ_{ik} and the functional prototypes $\phi_k(t)$ are assumed nonnegative, so that $\mu_i(t)$ satisfies the nonnegativity constraint for the mean of the Poisson distribution. This generative model lacks an overall trajectory $\mu(t)$ shared across subjects since it would function as a floor, rather than a mean, of the curves.

Expressing this model in matrix form, to reflect the discrete nature of the observed data, and approximating the functional prototypes with a spline basis expansion in terms of K_θ spline basis functions, we can rewrite equation (3.1) as $\boldsymbol{\mu}_i = \sum_{k=1}^{K} \xi_{ik} \boldsymbol{\Theta}_i \boldsymbol{\phi}_k$. Here $\boldsymbol{\Theta}_i$ is a $J_i \times K_\theta$ matrix of spline basis functions and $\boldsymbol{\phi}_k$ are spline coefficient vectors for the functional prototypes. We use B-spline basis functions, as they are nonnegative and so ensure the nonnegativity of the $\boldsymbol{\mu}_i$. During estimation, we set the number K_θ of spline basis functions to a large number and then encourage smoothness with a second derivative penalty on the spline coefficient functions. Letting $\mu_i(t_{ij})$ be the jth element of $\boldsymbol{\mu}_i$, the negative log penalized likelihood we minimize is

$$-\log\left\{\prod_{i=1}^{I}\prod_{j=1}^{J_i}\frac{\mu_i(t_{ij})^{Y_i(t_j)}e^{-\mu_i(t_j)}}{Y_i(t_j)}\right\} + \lambda\sum_{k=1}^{K}\boldsymbol{\phi}_k^T\boldsymbol{D}\boldsymbol{\phi}_k, \tag{3.2}$$

where $\boldsymbol{D}$ is the matrix that penalizes the second derivative of the spline coefficient functions.

We minimize this negative penalized log likelihood using the method of alternating minimization [Udell *et al.*, 2016]. This is a method that alternates between estimating the spline coefficient vectors, conditioning on the current values of the scores, and estimating the scores, conditioning on the current values of the spline coefficient vectors. Therefore we alternate between conditioning on the current values of the ξ_{ik} and estimating the $\boldsymbol{\phi}_k$, and vice versa. As discussed later, we select the value of λ using cross-validation. We use random nonnegative starting values for the scores.

Each of the two sub-problems that arises during alternating minimization amounts to fitting a generalized linear model with a nonnegativity constraint and, in the case of the model for the spline coefficient vectors, a second derivative penalty. To see this, assume for simplicity of exposition that all the trajectories are observed at the same set of times $t_1, t_2, \ldots, t_J$, so that all trajectories share a common $J \times K_\theta$ matrix $\boldsymbol{\Theta}$ of spline basis functions. Then, stacking all the observations in an $J \times I$ matrix $\boldsymbol{Y}$, our model for the observations (leaving out the penalty) can be expressed as $\boldsymbol{Y} \sim \text{Pois}(\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T)$, where $\boldsymbol{\Phi}$ is a $K_\theta \times K$ matrix of spline coefficient vectors and $\boldsymbol{\Xi}$ is an $n \times K$ matrix of scores. Combining all the spline coefficients into one matrix, this model can equivalently be expressed as $\boldsymbol{Y} \sim \text{Pois}(\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T)$, where 1_n is an n-vector of 1's. Using the matrix identity $\text{vec}\,(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\text{vec}\,(\boldsymbol{B})$, we can rewrite this expression as $\text{vec}\,(\boldsymbol{Y}) \sim \text{Pois}\{(\boldsymbol{\Xi} \otimes \boldsymbol{\Theta})\text{vec}\,(\boldsymbol{\Phi})\}$.

Using estimates of the scores and a fixed value of λ, $\boldsymbol{\Phi}$ in this last expression can be estimated by fitting a regularized Poisson generalized linear model with a nonnegativity constraint on the coefficients, using $\text{vec}(\boldsymbol{Y})$ as the responses, $\boldsymbol{\Xi} \otimes \boldsymbol{\Theta}$ as the model matrix and $\lambda \sum_{k=1}^{K} \boldsymbol{\phi}_k^T \boldsymbol{D} \boldsymbol{\phi}_k$ as the quadratic penalty. We do this using an implementation in the `nloptr` package of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, a quasi-Newton method that minimizes a local quadratic approximation to the objective function, that allows for box constraints [Byrd *et al.*, 1995].

To estimate the vector of scores $\boldsymbol{\xi}_i$ for the ith trajectory, we first write the model for the observations $\boldsymbol{Y}_i$ for the ith trajectory (again omitting the penalty) as $\boldsymbol{Y}_i \sim \text{Pois}(\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\xi}_i^T)$. Using estimates of $\boldsymbol{\Phi}$, $\boldsymbol{\xi}_i$ can also be estimated using a regularized Poisson linear model with a nonnegativity constraint, using $\text{vec}(Y_i)$ as the vector of responses and $\boldsymbol{\Theta}\boldsymbol{\Phi}$ as the model matrix. We fit this model using the `NNLM` package.

We use five-fold cross validation to select the best λ from a pre-defined sequence of values. Each fold is a group of approximately one-fifth of the curves. For each fold and value of λ, we fit our model using the other curves, and use the spline coefficients estimated using these curves to estimate scores, and thus fitted values, for the held-out curves. The criterion we use to select the optimal λ is the mean value of the Poisson likelihood of the held-out curves, given the predictions, over the five folds.

3.3 Generalized functional principal components analysis

In sections 3.4 and 3.5 we compare NARFD to generalized functional principal components analysis (GFPCA). This is a model that has been previously developed in the literature, which we estimate using a method similar to that we use for estimating NARFD models. Here we briefly describe the GFPCA model and how we estimate it using alternating minimization.

In GFPCA for Poisson data with a logarithmic link function, the following generative process for the trajectories is assumed:

$$\mu_i(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t)$$

$$Y_i(t) \sim \text{Pois}(\exp(\mu_i(t))),$$

where $\mu(t)$ is a mean trajectory shared across subjects and all other terms are as defined previously, although we refer to the $\phi_k(t)$ here as FPCs rather than as functional prototypes.

Several estimation strategies for this model have been proposed. Goldsmith *et al.* [2015] fit a Bayesian form of this model using the Hamiltonian Monte Carlo sampler implemented in `Stan` [Hoffman and Gelman, 2011]. van der Linde [2009] proposed a computationally efficient variational Bayes algorithm estimation method to fit a Bayesian form of this model. Hall *et al.* [2008] fit a similar model, one that assumes that deviations around the mean are small, by estimating the mean and covariance using observed data and then obtaining latent mean and basis functions by inverting a linear approximation to the logarithmic link function.

As the full Bayesian treatment in Goldsmith *et al.* [2015] is relatively slow, van der Linde [2009] did not make code implementing her method publicly available, and we found the method of Hall *et al.* [2008] to be numerically unstable (see Appendix Figures B.8 and B.9; this may occur because of violations of the assumption of this method that deviations around the mean are small), we have implemented an alternating minimization algorithm for estimating GFPCA. This involves a few modifications from the alternating minimization algorithm we use for estimating NARFD models. First, in addition to principal component spline coefficient vectors $\boldsymbol{\phi}_k$, we also estimate $\boldsymbol{\beta}$, a spline coefficient vector that estimates the mean function $\mu(t)$. Second, in lieu of a regularized Poisson generalized linear model with a nonnegativity constraint, we use a regularized Poisson generalized linear model with a logarithmic link. To estimate the mean and principal component coefficient vectors, both of which we penalize with the same second derivative penalty, we use the `mgcv` package [Wood, 2011]. To estimate the scores, we use a standard generalized linear models routine, including the estimated mean function as an offset. Third, after estimation of the FPCs in each iteration, we orthogonalize the FPCs using the singular value decomposition. This avoids degenerate solutions characterized by estimated FPCs that are multiples of each

other, and also improves interpretability of the estimated FPCs. After estimation of the scores in each iteration, we center the scores for each FPC around 0, and add the appropriate multiple of that FPC to the estimated mean.

3.4 Simulations

To simulate non-negative count data, we use the model $Y_i(t_j) \sim \text{Pois}[\mu(t) + h\{\xi_{i1}\phi_1(t_j) + \xi_{i2}\phi_2(t_j)\}]$, where h is either the identity function or the exponential function. In our first simulation scenario (which we call Scenario I), we assume data are generated using the generative model appropriate for modeling with NARFD. Here h is the identity function, $\mu(t) = 0$, $\phi_1(t)$ is the function $\sin(2s) + 1$ and $\phi_2(t)$ is the function $\cos(s) + 1$. The scores ξ_{i1} and ξ_{i2} are the squares of random variables generated from normal distributions with standard deviations equal to 4 and 3, respectively. The t_j for each j are a sequence of length 50 equally spaced over the domain $[0, 2\pi]$, so that $J_i = 50$ for all i.

We also simulate data under a second scenario (Scenario II), where we assume that data are generated using the generative model appropriate for modeling with GFPCA. We simulate data under this second scenario both to confirm that our implementation of GFPCA accurately estimates GFPCA models, and also to compare how the GFPCA and NARFD models perform under model misspecification (where, for example, data is generated with the GFPCA generative model, and a NARFD model is used for estimation). For this second scenario, h is the exponential function, $\mu(t) = 3$, $\phi_1(t)$ is the function $\sin(2s)$, and $\phi_2(t)$ is the function $\cos(s)$. The scores ξ_{i1} and ξ_{i2} are generated from normal distributions with standard deviations equal to 1.5 and 1.0, respectively.

Figure 3.2 shows NARFD estimates of $\phi_1(t)$ and $\phi_2(t)$ in Scenario I, for $I \in \{50, 200, 400\}$. Estimation quality improves as I increases. Appendix Figure B.1 shows GFPCA estimates of $\phi_1(t)$ and $\phi_2(t)$ in Scenario II. Again, estimation quality improves as I increases, and the correct FPCs are recovered. Appendix Figures B.2 and B.3 show how integrated squared errors of estimation improve for both NARFD and GFPCA when I increases, when the generative model matches the estimation method used.

To examine how the two methods perform under model misspecification, we simulated

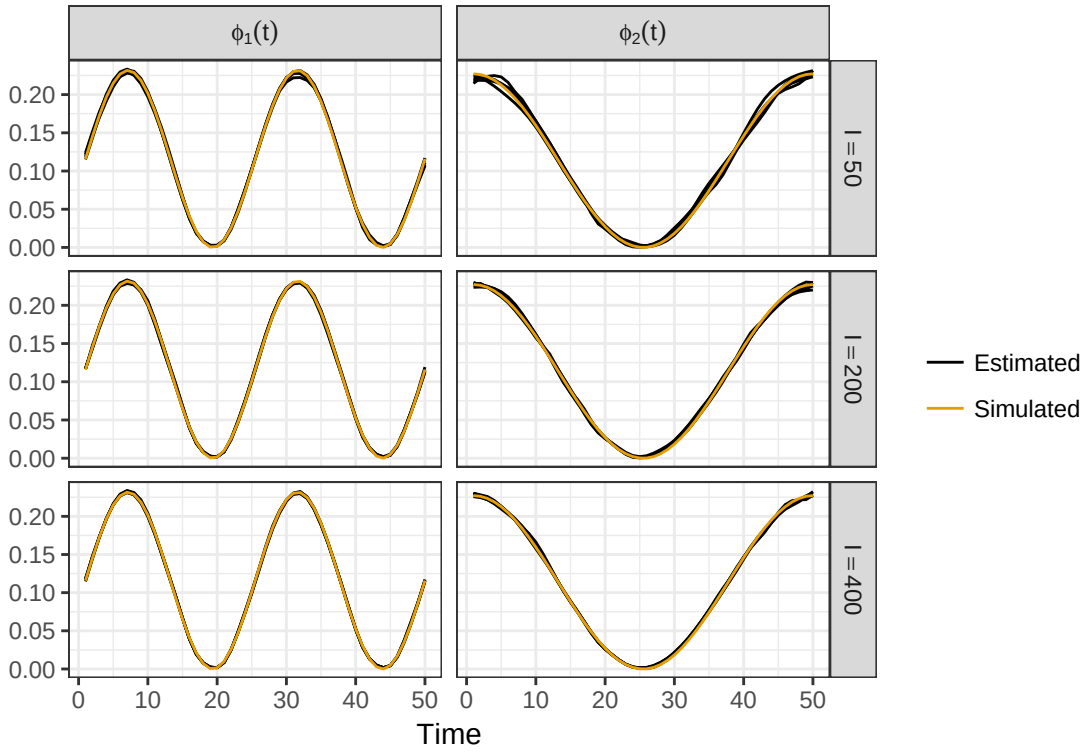

Figure 3.2: Simulated FPCs and NARFD estimates for Scenario I, for different numbers of curves per simulation replicate. Each simulation was replicated 5 times.

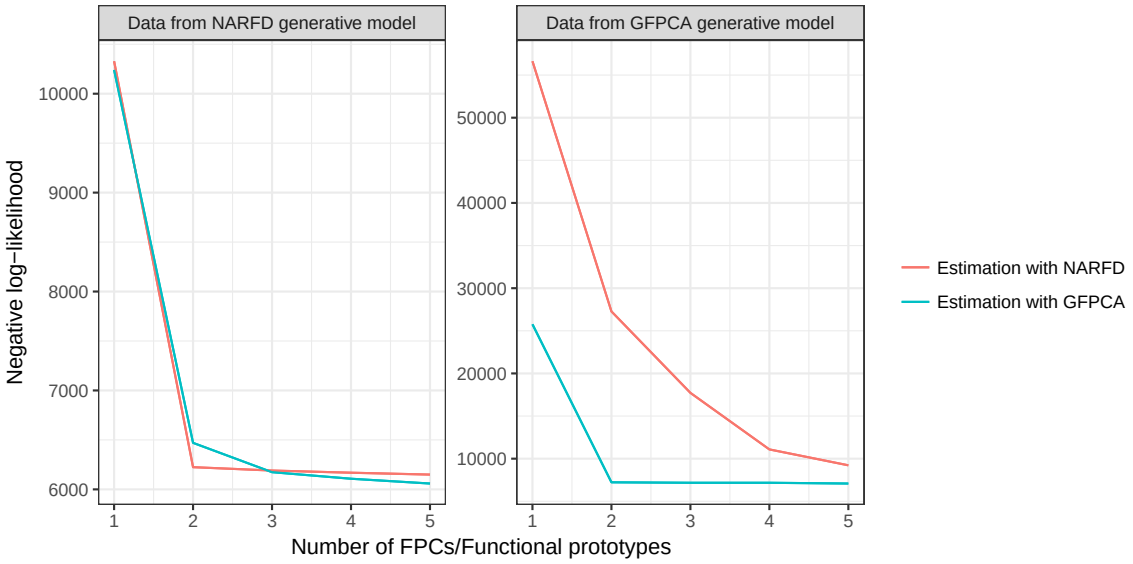

Figure 3.3: Negative Poisson log-likelihood of data generated using the NARFD generative model and fitted using NARFD and GFPCA, left, and of data generated using the GFPCA generative model and fitted using NARFD and GFPCA, right. Here $I = 50$ and $K_\theta = 25$.

data using Scenario II and estimated functional prototypes using NARFD, and simulated data using Scenario I and estimated FPCs using GFPCA. Reconstruction of the curves is better when the estimation method matches the generative model, although reconstruction error decreases with additional functional prototypes/FPCs, more quickly with GFPCA than with NARFD, when the estimation method does not match the data generating model (see Appendix Figure 3.3). As expected, given the different generative models and constraints applicable to the two methods, neither method can accurately recover FPCs/functional prototypes used in simulating data using the other method (see Appendix Figure B.10).

The Appendix also includes results showing the effect of changing the number of basis functions (see Appendix Figures B.4 and B.5) and estimating more FPCs/functional prototypes than are used in simulation (see Appendix Figures B.6 and B.7). Due to its nonnnegativity constraint, NARFD requires many basis functions to accurately represent quickly varying curves. Figures B.8 and B.9 also show results, for Scenario II, using the method of Hall *et al.* [2008], for comparison with our implementation of GFPCA.

For Scenario I with 400 curves, one run of NARFD at the λ selected by cross-validation took about 80 seconds. For Scenario II with 400 curves, one run of GFPCA at the λ selected by cross-validation took about 90 seconds.

3.5 Results

We apply NARFD to data from the Baltimore Longitudinal Study on Aging (BLSA) [Schrack *et al.*, 2014], a study of human aging with healthy participants. We also compare with our implementation of GFPCA. The sample we consider in this chapter consists of 631 men and women who wore the Actiheart, a combined heart rate and physical activity monitor placed on the chest [Brage *et al.*, 2006]. Subjects were asked to wear the device except when showering, bathing or swimming. Physical activity was measured in activity counts per minute, a cumulative summary of acceleration detected by the device within 1-minute monitoring epochs [Bai *et al.*, 2014].

Subjects in our BLSA sample have between 1 and 26 days of monitoring data. To obtain consistent activity profiles, we selected the 592 subjects with at least 5 days of data, added observations across the first 5 days of observations for those subjects, and then combined the activity data into 10 minute intervals. This yields 144 observations for each subject, combining activity across 5 days of monitoring. Given the periodic nature of the data, we use a periodic B-spline basis, with 25 basis functions.

To compare the methods, we estimated FPCs/functional prototypes with each method using from 1 to 12 FPCs/functional prototypes. We used 50 subjects to estimate FPCs and then estimated scores for the remaining held-out subjects. Figure 3.4 shows the negative Poisson log-likelihood with respect to curves of these held-out subjects for NARFD and GFPCA as a function of the number of FPCs/functional prototypes used in the decomposition. NARFD is less parsimonious than BLSA in explaining the variation in the BLSA data. Figure 3.5 shows the FPCs estimated using each method, for a model fit with five FPCs using all the 592 curves. FPCs/functional prototypes are ordered for GFPCA by the standard deviation of the corresponding scores, after normalization of the FPCs, and for NARFD by the sum of the contribution of the functional prototypes to the curve recon-

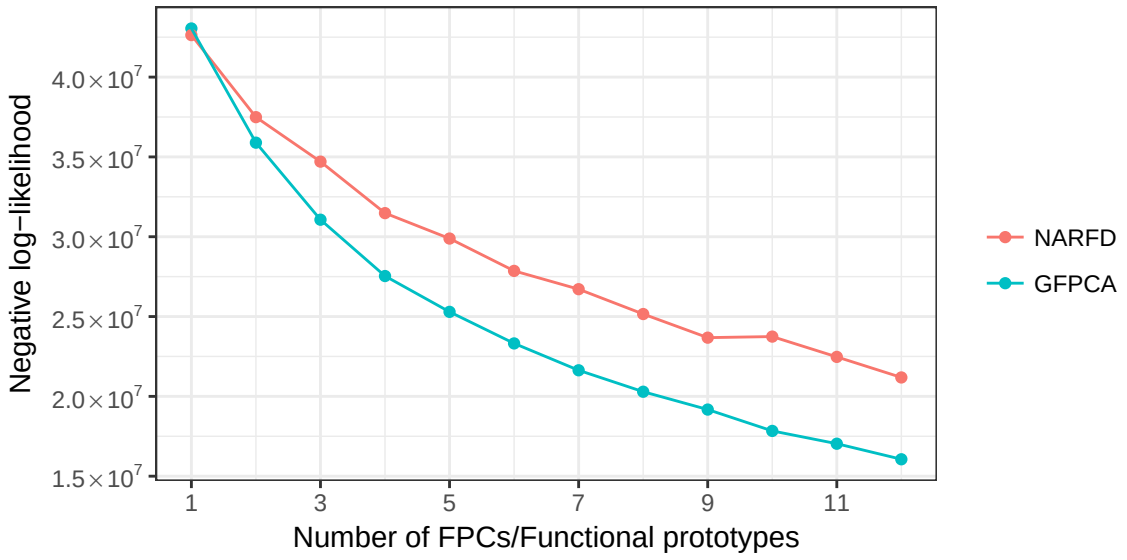

Figure 3.4: Negative Poisson log-likelihood for held-out curves from BLSA data for NARFD and GFPCA, decomposed using 1 through 12 FPCs/functional prototypes estimated using 50 curves from BLSA data.

structions. The functional prototypes estimated using NARFD are simpler, sparser, and easier to interpret than the FPCs estimated using GFPCA, each directly corresponding to a burst of activity at particular times during the day.

Figure 3.6 shows the reconstruction of a curve using NARFD and GFPCA. NARFD has the property that the contributions from each prototype are additive, so that the final curves incorporate the contributions from each FPC without cancellation. On the other hand, contributions from FPCs for GFPCA may cancel out, since the FPCs estimated using GFPCA are multiplicative, and the contribution of a FPC to the overall activity profile depends not only on the score for that FPC but also on the scores for the other FPCs.

3.6 Discussion

We have presented NARFD, a novel decomposition of non-negative functional count data which enables the study of patterns of variation across subjects in a highly interpretable

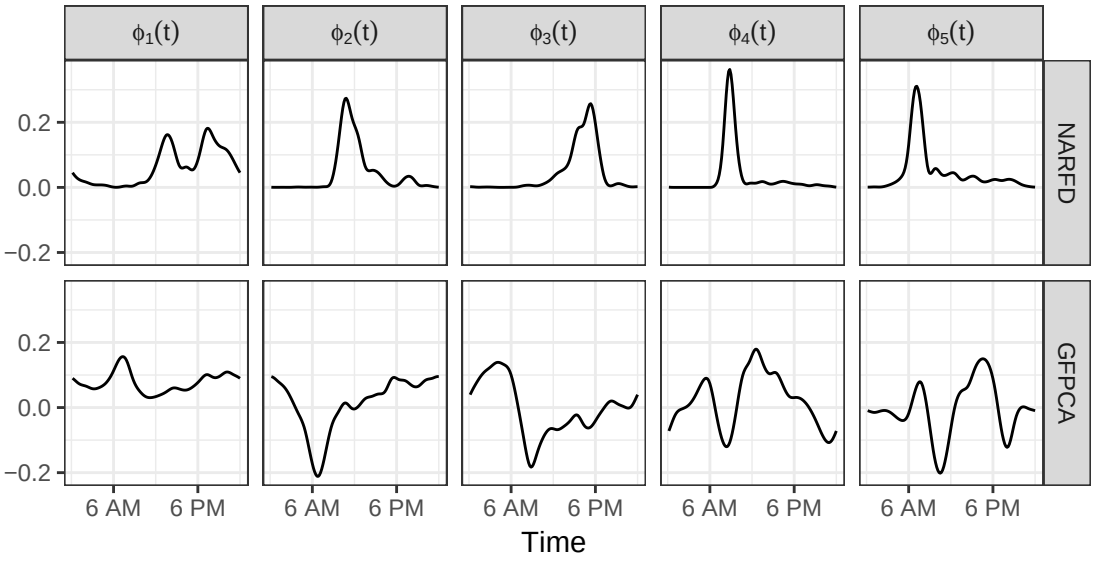

Figure 3.5: First five estimated FPCs/functional prototypes for BLSA data. GFPCA FPCs are shown on the scale on which they are estimated (prior to exponentiation).

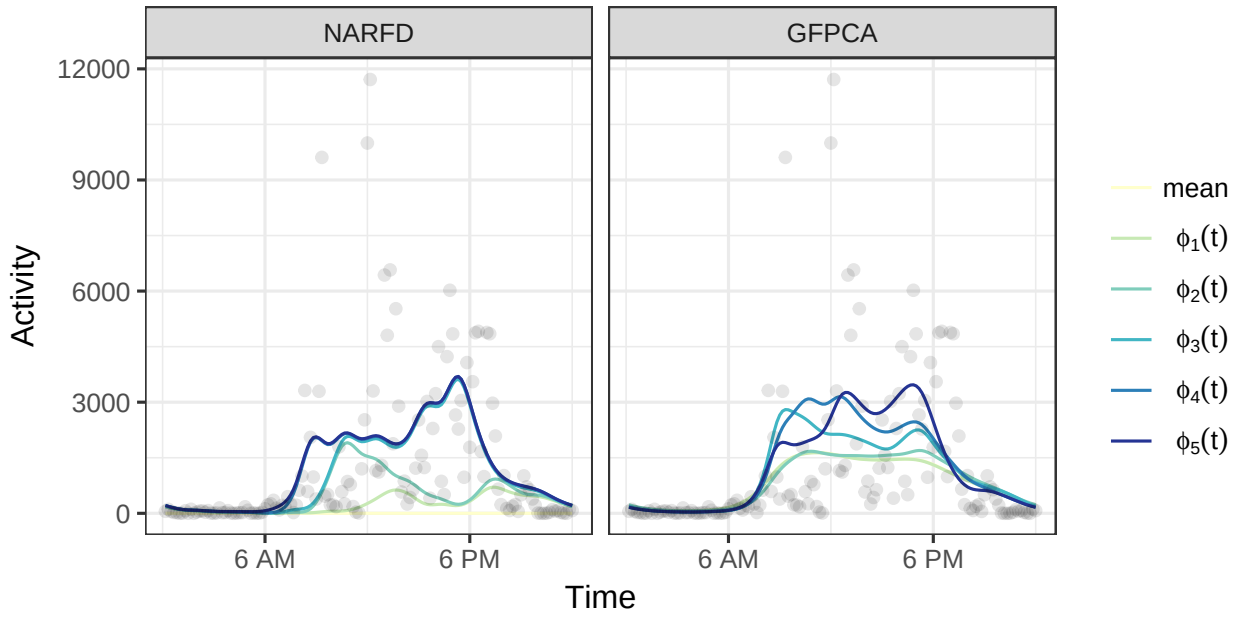

Figure 3.6: Reconstruction of a subject's data using 5 FPCs/functional prototpyes, with NARFD and GFPCA. Activity counts are shown in light dots, and cumulative contributions of the mean and the FPCs/functional prototypes are shown as lines.

manner. Applying these methods to our motivating dataset, we have extracted functional components which show clear peaks of activity at various times during the day. The accompanying scores can be used to classify subjects, and can be used as outcome variables to investigate the relationship between covariates and activity at different times of the day.

We have also presented a novel algorithm for fitting GFPCA models for count data with a logarithmic link, using alternating minimization. Both of our methods can accurately recover correct modes of variation, as demonstrated in our simulation studies.

Further work in this area could investigate ways to have a separate smoothness penalty for each FPC/functional prototype. This would require the development of a new procedure for smoothness parameter selection, as our cross-validation procedure would become computationally infeasible with more than a few smoothness parameters. Further work could also incorporate random effects, so that, as in [Goldsmith *et al.*, 2015], patterns of activity within subjects, in addition to patterns of activity across subjects, could be analyzed.

Part II

Methods in functional genomics

Chapter 4

FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation

4.1 Introduction

Understanding the functional consequences of noncoding genetic variation is one of the most important problems in human genetics. Comparative genomics studies suggest that most of the mammalian conserved and recently adapted regions consist of noncoding elements [Lindblad-Toh *et al.*, 2011; Khurana *et al.*, 2013; ENCODE Project Consortium, 2012]. Furthermore, most of the loci identified in genome-wide association studies fall in noncoding regions and are likely to be involved in gene regulation in a cell type and tissue specific manner [Altshuler *et al.*, 2008]. Noncoding variants are also known to play an important role in cancer. Somatic variants in noncoding regions can act as drivers of tumor progression and germline noncoding variants can act as risk alleles [Khurana *et al.*, 2016]. Thus, improved understanding of tissue-specific functional effects of noncoding variants will have implications for multiple diseases and traits.

Prediction of the functional effects of genetic variation is difficult for several reasons. To begin with, there is no single definition of function. As discussed in Kellis *et al.* [2014] there are several possible definitions, depending on whether one considers genetic, evolutionary conservation or biochemical perspectives. These different approaches each have limitations and vary substantially with respect to the specific regions of the human genome that they predict to be functional. In particular, the genetic approach, based on experimental evaluation of the phenotypic consequence of a sequence alteration (e.g. by measuring the impact of individual alleles on gene expression in a particular context), is currently laborious, has modest throughput and may miss elements that lead to phenotypic effects manifest only in rare cells or specific contexts. The evolutionary approach relies on accurate multispecies alignment which makes it challenging to identify certain functional elements, such as distal regulatory elements known to evolve rapidly, although recently several approaches have been developed for primate- or even human-specific elements [Petrovski *et al.*, 2013]. An additional limitation of the evolutionary approach is that it is not sensitive to tissue and cell type. Finally, the biochemical approach adopted by projects such as ENCODE [ENCODE Project Consortium, 2012] and Roadmap Epigenomics [Roadmap Epigenomics Consortium, 2015], although helpful in identifying potentially regulatory elements in specific contexts, does not provide definitive proof of function since the observed biochemical signatures can occur stochastically and in general are not completely correlated with function. Besides the difficulty in precisely defining function, a challenge is that the use of functional genomics features from ENCODE and Roadmap Epigenomics (e.g. ChIP-seq and DNase I hypersensitive sites signals) are mostly useful for predicting the effects of variants in cis-regulatory elements, such as promoters, enhancers, silencers and insulators. Other classes of functional variants, for example those with effects on post-transcriptional regulation by alteration of RNA secondary structure or RNA-protein interactions, would be missed by these features.

Recently, several computational approaches have been proposed to predict functional effects of genetic variation in noncoding regions of the genome based on epigenetic and evolutionary conservation features [Khurana *et al.*, 2013; Kircher *et al.*, 2014; Fu *et al.*, 2014; Ionita-Laza *et al.*, 2016; Quang *et al.*, 2015; Huang *et al.*, 2017]. These predictions are at the organism level and are not specific to particular cell types or tissues. Here we are interested

in predicting functional effects of genetic variants in specific cell types and tissues using epigenetic features and chromatin accessibility measurements. The ENCODE Project and the Roadmap Epigenomics Project have profiled various epigenetic features, including histone modifications and chromatin accessibility, genome-wide in more than a hundred different cell types and tissues. Histone modifications are chemical modifications of the DNA-binding histone proteins that influence transcription as well as other DNA processes. Particular histone modifications have characteristic genomic distributions [Bannister and Kouzarides, 2011]. For example, trimethylation of histone H3 lysine 4 (H3K4me3) is associated with promoter regions, monomethylation of histone H3 lysine 4 (H3K4me1) is associated with enhancer regions, and acetylation of histone H3 lysine 27 (H3K27ac) and of histone H3 lysine 9 (H3K9ac) is associated with increased activation of enhancer and promoter regions [Roadmap Epigenomics Consortium, 2015]. Repressive marks include trimethylation of histone H3 lysine 27 (H3K27me3) and trimethylation of histone H3 lysine 9 (H3K9me3), both associated with inactive promoters of protein-coding genes; H3K27me3 is found in facultatively repressed genes by Polycomb-group factors, while H3K9me3 is found in heterochromatin regions corresponding to constitutively repressed genes [Friedman and Rando, 2015].

Several unsupervised approaches exist for the integration of these epigenetic features in specific cell types and tissues. Such integrative approaches reflect the belief that epigenetic features interact with one another to control gene expression. One class of methods attempts to segment the genome into non-overlapping segments, representing major patterns of chromatin marks, and labels these segments using a small set of labels such as active transcription start site, enhancer, strong transcription, weak transcription, quiescent etc. This class includes methods such as ChromHMM [Roadmap Epigenomics Consortium, 2015; Ernst and Kellis, 2012, 2015] and Segway [Hoffman *et al.*, 2012], based on Hidden Markov Models (HMMs) and Dynamic Bayesian Networks respectively. ChromHMM is based on complete pooling of data from multiple tissues and fitting a single model to this superdataset, while Segway is based on fitting separate models to data from each tissue (no pooling). Various extensions of these early segmentation approaches have been proposed. Several approaches have focused on better modeling the read count data from the under-

lying assays using Poisson-lognormal and negative multinomial distributions [Zacher *et al.*, 2017; Mammana and Chung, 2015], while others have focused on better modeling of the correlations among related cell types and tissues [Biesinger *et al.*, 2013; Zhang *et al.*, 2016; Zhang and Hardison, 2017]. Yet another approach attempts to improve the HMM parameter estimation procedure in ChromHMM by replacing the EM algorithm with a spectral learning procedure [Song and Chen, 2015]. Another class of methods focuses exclusively on predicting functional effects of variants, rather than segmenting the genome as discussed above. A recent method in this class, GenoSkyline [Lu *et al.*, 2016], is based on fitting a two-component mixture model of multivariate Bernoulli distributions to epigenetic data for each tissue separately, and then computing a posterior probability for each variant to be in the functional class. Recently, several supervised approaches have been proposed as well, including deltaSVM [Lee *et al.*, 2015] and cepip [Li *et al.*, 2017]. While supervised approaches can be more efficient than unsupervised ones when high-quality, unbiased labeled data are available for training, unsupervised approaches as proposed here can provide more robust, less biased functional predictions across large number of tissues and cell types when such unbiased labeled data is scarce, as is the case now.

We introduce here a new integrated functional score that combines different epigenetic features in specific cell types and tissues. Our model is based on the latent Dirichlet allocation (LDA) model [Blei *et al.*, 2003], a generative probabilistic model often used in the topic modeling literature, that allows joint modeling of data from multiple cell types and tissues. In our context, the latent functional classes correspond to latent topics in the topic modeling setting, the various tissues correspond to different documents, while the tissue specific position scores correspond to words in a document. The proposed LDA model has several advantages. First, our method makes no distributional assumptions on the data, allowing us to avoid various data transformations employed by other approaches (such as binary peak calling/dichotomization), and facilitating the integration of annotation data on the original scale (e.g. quantitative, binary etc.). Second, because the model is fit jointly to data from multiple cell types and tissues, cross-tissue comparisons are meaningful. Third, we show that our method outperforms other methods in labeling positions in the genome as functional in a particular tissue, or not.

4.2 Methods

4.2.1 LDA model for functional annotation

We propose an application of the latent Dirichlet allocation (LDA) model [Blei *et al.*, 2003], a generative probabilistic model, in the setting of functional genomics annotation with the goal of computing posterior probabilities for positions to belong to different functional classes in any given tissue.

The position scores in each tissue are modeled as a mixture over latent functional classes. In the mixture distribution, we assume that the mixture components are shared across all the tissues, while the mixture proportions for the different functional classes can vary from tissue to tissue. Since our primary goal is to provide a functional score (as opposed to a functional element annotation) we focus on integrating four activating histone modifications (i.e. H3K4me1, H3K4me3, H3K9ac, H3K27ac) and DNase. For the four activating histone modifications data, we compute "valley" scores, motivated by previous work showing that within regions of high histone acetylation, local minima (or valleys) are strongly associated with transcription factor binding sites. We fit the LDA model with nine functional classes to these data, and compute for each position its posterior probability to belong to a functional class in a specific tissue. We define the functional score at a position as the sum of posterior probabilities for the designated 'active enhancer' and 'active promoter' classes.

We now present a detailed description of our model and how we estimate it. Let us assume that we have a set of m genomic positions in the training set, together with a set of k functional annotations. For each position i, we have k tissue-specific functional scores: $\mathbf{X_i} = (X_{i1}, \ldots, X_{ik})$. Let $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_m})$ be the set of (continuous) functional scores for all the positions. These scores are epigenetic features (histone modifications and DNase hypersensitivity) from ENCODE and Roadmap Epigenomics across a varied set of tissues and cell types. Let l be the number of tissues, and m_j be the number of positions with tissue j annotations in the training set ($m = \sum_{j=1}^{l} m_j$). For each position $i \leq m$ in the training set we denote by t_i the corresponding tissue (i.e. the annotations corresponding to this variant are for tissue t_i). For each tissue, the positions' scores are represented as a mixture over latent functional classes, where each functional class is characterized by a

distribution over position scores. In what follows, for ease of presentation, we assume only two latent functional classes, but the number of classes can be chosen to be greater than two. We let $\mathbf{C} = (C_1, \ldots, C_m)$ denote the set of indicator variables for all the positions, where $C_i = 1$ if position i belongs to the first functional class and $C_i = 0$ otherwise. We are not able to observe $\mathbf{C}$.

Let $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ be the hyperparameter vector with $\alpha_0, \alpha_1 > 0$. Here we assume $\boldsymbol{\alpha}$ is a vector of 1s throughout (a uniform prior). We assume the functional annotation data has been generated from the following generative model:

1. For each tissue j, choose $(1 - \pi_j, \pi_j) \sim \text{Dir}(\alpha_0, \alpha_1)$.

2. Given π_j, for each position i with $t_i = j$ choose a class $C_i \sim \text{Bern}(\pi_j)$.

3. Given $C_1, \ldots, C_m$, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are independently generated with each $\boldsymbol{X}_i$ being generated from the appropriate multivariate distribution: F_1 if $C_i = 1$, and F_0 otherwise.

Here $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_l)$ and $\boldsymbol{C}$ are latent variables. We want to calculate the posterior probability for each position i to be in the first functional class:

$$w_i = P(C_i = 1 \,|\, \boldsymbol{X}_i, \boldsymbol{\alpha}),$$

and the densities f_0 and f_1. For a given tissue the conditional density of $(\boldsymbol{\pi}, \boldsymbol{C})$ given $\boldsymbol{X}$ and $\boldsymbol{\alpha}$ is:

$$p(\pi, \boldsymbol{C} | \boldsymbol{X}, \boldsymbol{\alpha}) = \frac{p(\pi, \boldsymbol{C}, \boldsymbol{X} | \boldsymbol{\alpha})}{p(\boldsymbol{X} | \boldsymbol{\alpha})}.$$

For the numerator we have:

$$p(\pi, \boldsymbol{C}, \boldsymbol{X} | \alpha) = p(\pi | \boldsymbol{\alpha}) \prod_{i=1}^{m} p(C_i | \pi) p(\boldsymbol{X}_i | C_i).$$

This is easy to compute. However the denominator is not. For the denominator we have:

$$p(\boldsymbol{X} | \boldsymbol{\alpha}) = \int p(\pi | \boldsymbol{\alpha}) \left(\prod_{i=1}^{m} \sum_{C_i} p(C_i | \pi) p(\boldsymbol{X}_i | C_i) \right) d\pi.$$

There are 2^m terms in the summation so this is difficult to compute for moderately large m. We propose instead to use a variational approach as described in Blei *et al.* [2003]. In the variational inference approach we first introduce a family of distributions $\{q(\cdot, \cdot | \boldsymbol{a}, \boldsymbol{w})\}$

over the latent variables (π, C_m) with its own variational parameters $\boldsymbol{a} = (a_0, a_1)$ and $\boldsymbol{w}$ (these are tissue specific parameters).

Then

$$q(\pi, C_m|\boldsymbol{a}, \boldsymbol{w}) = q(\pi|\boldsymbol{a}) \prod_{i=1}^{m} q(C_i|w_i),$$

where $q(\pi|\boldsymbol{a})$ is the density of $\text{Dir}(\boldsymbol{a})$ and $q(C_i|w_i)$ is the probability mass function of $\text{Bern}(w_i)$ for $i = 1 \dots m$.

Using Jensen's inequality we have:

$$\begin{aligned}
\log p(\boldsymbol{X}|\boldsymbol{\alpha}) &= \log \int \sum_{\boldsymbol{C}} p(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{X}|\boldsymbol{\alpha})\, d\boldsymbol{\pi} \\
&= \log \int \sum_{\boldsymbol{C}} \frac{p(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{X}|\alpha)}{q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w})} q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w})\, d\pi \\
&\geq \int \sum_{\boldsymbol{C}} q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w}) \log p(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{X}|\boldsymbol{\alpha})\, d\boldsymbol{\pi} - \int \sum_{\boldsymbol{C}} q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w}) \log q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w})\, d\pi \\
&= E_q \log p(\boldsymbol{\pi}, \boldsymbol{C}, \boldsymbol{X}|\alpha) - E_q \log q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w}) = L(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{\alpha}).
\end{aligned}$$

Note that $L(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{\alpha})$ is a lower bound on the log likelihood. So instead of maximizing the log likelihood directly we maximize this lower bound with respect to the variational parameters $\boldsymbol{a}$ and $\boldsymbol{w}$. It can be shown that $\log p(\boldsymbol{X}|\boldsymbol{\alpha}) - L(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{\alpha})$ is the Kullback-Leibler (KL) divergence between the true posterior $p(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{\alpha}, \boldsymbol{X})$ and the variational posterior $q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w})$ with respect to $q(\boldsymbol{\pi}, \boldsymbol{C}|\boldsymbol{a}, \boldsymbol{w})$. Therefore by maximizing $L(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{\alpha})$ with respect to $\boldsymbol{a}$ and $\boldsymbol{w}$, we minimize the KL divergence between the variational posterior probability and the true posterior probability. Then we can estimate $P(C_i = 1|\boldsymbol{\alpha}, \boldsymbol{X})$ by w_i for each variant i. Below we describe the variational inference algorithm.

Variational Inference Algorithm Assume the initial state $\big(w_1, \dots, w_m, f_0, f_1\big)$. The algorithm proceeds as follows:

Step 1. *(Kernel Density Estimation)*

Fit a multivariate kernel density estimate for each annotation and component separately: f_{0s}^{new} and f_{1s}^{new} for each annotation $s = 1, \dots, k$, weighting variants by component membership probability. Specifically, for any $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^k$ and

$s = 1, \ldots, k$, we let

$$f_{0s}^{\text{new}}(x_s) = \frac{\sum_{i=1}^{m}(1-w_i)\mathrm{K}_{h_s}(x_s - X_{is})}{\sum_{i=1}^{m}(1-w_i)},$$

and

$$f_{1s}^{\text{new}}(x_s) = \frac{\sum_{i=1}^{m} w_i \mathrm{K}_{h_s}(x_s - X_{is})}{\sum_{i=1}^{m} w_i}.$$

The scaled kernel $\mathrm{K}_{h_s}(a) = \frac{1}{h_s}K(\frac{a}{h_s})$, where $K(\cdot)$ is taken to be the probability density function of a standard normal, and the bandwidth parameter h_s is chosen to be

$$h_s = 0.9 \min\{\mathrm{SD}_s, \mathrm{IQR}_s/1.34\} m^{-1/5}$$

according to a rule of thumb due to Silverman Silverman [1986], where SD_s and IQR_s are the standard deviation and interquartile range of annotation s, respectively. Then

$$f_0^{\text{new}}(\boldsymbol{m}x) = \prod_{s=1}^{k} f_{0s}^{\text{new}}(x_s), \quad \text{and} \quad f_1^{\text{new}}(\boldsymbol{m}x) = \prod_{s=1}^{k} f_{1s}^{\text{new}}(\boldsymbol{m}x).$$

Step 2. *(Variational Step)*

For each tissue j, we obtain w_i for all variants i with $t_i = j$ and (a_0^j, a_1^j) by maximizing the lower bound on the marginal likelihood of $\boldsymbol{X}$, i.e. $L(\boldsymbol{a}, \boldsymbol{w}|\boldsymbol{\alpha})$, with respect to $\boldsymbol{a}$ and $\boldsymbol{w}$.

This results in the following iterative algorithm:

$$w_i = \frac{f_1(\boldsymbol{X}_i)\, \exp(\Psi(a_1^j)))}{f_1(\boldsymbol{X}_i)\, \exp(\Psi(a_1^j)) + f_0(\boldsymbol{X}_i)\, \exp(\Psi(a_0^j))} \quad \text{for variants } i \text{ with } t_i = j,$$

$$a_0^j = \alpha_0 + \sum_{t_i=j}(1-w_i) \quad \text{and} \quad a_1^j = \alpha_1 + \sum_{t_i=j} w_i.$$

where $\Psi(x) = d\log\Gamma(x)/dx$ and $\Gamma(x)$ is the Gamma function. Additional details are available in Blei *et al.* [2003].

4.2.2 LDA implementation

We have implemented the above algorithm in an R package, FUNLDA. For training purposes, we select $4,000$ random positions in each of the 127 tissues. The positions are chosen among $9,254,335$ SNPs with minor allele count greater than 5 in European samples from

the 1000 Genomes project. We have also looked at other ways to select positions in the training set (e.g. randomly from across the entire genome, with enrichment near genes) and the results were similar, suggesting that our predictions are robust to the choice of positions used in the training sets. The number of outer iterations in the variational inference algorithm is 250 and the number of inner iterations is 200.

FUN-LDA is computed by fitting the LDA model with nine classes to and DNase hypersensitivity and valley scores for the four activating histone modifications (H3K4me1, H3K4me3, H3K9ac, H3K27ac). For the histone modifications and DNase we start with the negative log10 of the Poisson P-value of ChIP-seq or DNase counts relative to expected background counts, as output by ChromImpute [Ernst and Kellis, 2015]. The valley scores are computed as in Ramsey *et al.* [2010]: for every window of 25 bp, we calculate the maximum score for the two regions from -100 to -500 bp and from 100 to 500 bp. If the score at the window of 25 bp is less than 90% of the minimum of those two maxima, we set the value in that window to that minimum. Otherwise, we set the value in that 25 bp window to 0. For each variant, we get a set of nine posterior probabilities for the position to be in a specific functional class. To get a functional score, we sum the posterior probabilities for the active functional classes, namely 'active promoters' and 'active enhancers' (Figure 4.1), which we identified by looking for the characteristic histone modification and DNAse signatures associated with these active functional classes among the 9 classes inferred by our model.

4.3 Validation of our method

To assess the accuracy of the predictions of FUNLDA and compare with other methods, we conducted thorough comparisons using validation sets of variants that have been shown to have some evidence of a regulatory function. We use both tissue/cell type specific validation sets, and non tissue/cell type specific validation sets.

4.3.1 Tissue/cell type specific validation sets

We focus on several lists of variants with tissue/cell type specific functional evidence:

H3K27ac-V	H3K4me1-V	H3K4me3-V	H3K9ac-V	DNase	Size	Annotation
25.27	4.05	36.78	17.38	25.00	0.40%	ActivePromoters
2.99	2.80	1.02	0.93	4.33	1.59%	ActiveEnhancers
1.15	1.59	0.46	0.57	1.32	1.67%	WeakEnhancers
0.56	0.94	0.29	0.43	0.71	3.50%	NotFunctional
0.06	0.11	0.03	0.11	0.21	7.00%	NotFunctional
0.03	0.05	0.03	0.03	0.55	35.60%	NotFunctional
0.06	0.20	0.03	0.07	0.32	9.10%	NotFunctional
0.23	0.28	0.22	0.28	0.34	35.60%	NotFunctional
0.35	0.56	0.27	0.37	0.36	5.46%	NotFunctional

Figure 4.1: Heatmap showing classes inferred by FUN-LDA. The five left-most columns each show the average value of valley scores or the DNase hypersensitivity assay for positions assigned to the corresponding class, across all tissues. The sixth column indicates the percentage of positions assigned to each of the classes. The last column shows our assignment of function to the class. We sum the probability of being in the ActivePromoters and ActiveEnhancers rows to get the FUN-LDA score. The ActivePromoters state is characterized by high values of DNase and H3K4me3; the ActiveEnhancers state is characterized by high values of H3K4me1 and lower values of H3K4me3.

- confirmed regulatory variants from a multiplexed reporter assay in lymphoblastoid cell lines [Tewhey and others, 2016],
- regulatory motifs in 2,000 predicted human enhancers using a massively parallel reporter assay in two human cell lines, liver carcinoma (HepG2) and erythrocytic leukemia (K562) cell lines [Kheradpour *et al.*, 2013], and
- a collection of dsQTLs (DNase I sensitivity quantitative trait loci) in lymphoblastoid cell lines [Degner and others, 2012].

4.3.1.1 Confirmed regulatory variants (emVars) from a multiplexed reporter assay

In Tewhey and others [2016], the authors applied the massively parallel reporter assay (MPRA) to identify variants with effects on gene expression. In particular, they apply it to 32,373 variants from 3,642 cis-expression quantitative trait loci and control regions in lymphoblastoid cell lines (LCLs), and identify 842 variants showing differential expression between alleles, or emVars, expression-modulating variants. We use this set of 842 emVars as positive control variants. We paired each positive control with four variants tested using the MPRA where neither allele showed differential expression relative to the control, applying a threshold of 0.1 for the Bonferroni corrected p value. After removing from the list of positive and negative control variants those variants that we could not map to a genomic location using the Ensembl database (http://grch37.ensembl.org/index.html), there remained 693 positive control variants and 2,772 negative control variants.

We compute AUROC (the area under a receiver operating characteristic curve) values for several methods, including FUN-LDA, GenoSkyline, ChromHMM (25 state model), Segway, IDEAS and cepip (two versions: cepip_cell, and cepip_combined). For ChromHMM we partition the twenty-five states into two groups, 'functional' and 'non-functional', with the functional group consisting of 'TssA' (active TSS), 'PromU' (Promoter Upstream TSS), 'PromD1' (Promoter Downstream TSS 1), 'PromD2' (Promoter Downstream TSS 2), 'EnhA1' (Active Enhancer 1), 'EnhA2' (Active Enhancer 2), 'EnhAF' (Active Enhancer Flank). For each variant, the sum of ChromHMM posterior probabilities for the classes in the functional

group above is used to score the variant. Segway and IDEAS only provide a functional class assignment for each position, and we use these assignments to identify the functional variants. Results are shown in Table 4.1. As shown, FUN-LDA has higher AUROC (0.707) compared to existing tissue-specific functional prediction methods methods. FUN-LDA performs significantly better than the two binarized versions, but it does not outperform raw DNase (0.718).

4.3.1.2 Regulatory motifs in 2,000 predicted human enhancers using a massively parallel reporter assay

In Kheradpour *et al.* [2013], the authors use a massively parallel reporter assay to measure the transcriptional levels produced by targeted motif disruptions in $2,104$ candidate enhancers in two human cell lines, liver carcinoma (HepG2) and erythrocytic leukemia (K562) cell lines, providing one of the largest resource of experimentally validated enhancer manipulations in human cells. We use as positive control variants those variants where the p value comparing expression values for the sequence with the motif compared to sequences with scrambled versions of the motif was less than 0.05. We use as negative control variants those variants where this p value was greater than 0.1. After removing those variants whose genomic coordinates we could not resolve, there remained, for HepG2, 525 positive and $1,451$ negative control variants, and for K562, 342 positive and $1,578$ negative control variants. For all methods, we calculate the scores for these motifs by averaging across all bases in the motifs. As shown in Table 4.1, FUN-LDA has better accuracy than GenoSkyline, ChromHMM, IDEAS, Segway and cepip.

4.3.1.3 dsQTLs (DNase I sensitivity quantitative trait loci) in lymphoblastoid cell lines

We also utilized a collection of dsQTLs in human lymphoblastoid cell lines, originally identified using DNase I sequencing data from human lymphoblastoid cell lines [Degner and others, 2012]. In Lee *et al.* [2015] the authors further processed this list of dsQTLs and generated 579 dsQTLs (with p value $< 1 \times 10^{-5}$), and randomly selected as controls a larger set of common SNPs (minor allele frequency $> 5\%$) only from the top 5% of DNase

Table 4.1: Tissue/cell type specific functional predictions.

Dataset	Method	AUROC
emVars in Tewhey and others [2016], E116	FUN-LDA	0.707
	GenoSkyline	0.673
	ChromHMM	0.669
	Segway	0.622
	IDEAS	0.645
	DNase	0.718
	DNase-narrow	0.666
	DNase-gapped	0.659
	cepip_cell	0.653
	cepip_combined	0.642
Regulatory motifs in Kheradpour *et al.* [2013], E118/HepG2	FUN-LDA	0.691
	GenoSkyline	0.629
	ChromHMM	0.606
	Segway	0.618
	IDEAS	0.546
	DNase	0.719
	DNase-narrow	0.561
	DNase-gapped	0.550
	cepip_cell	0.592
	cepip_combined	0.641
Regulatory motifs in Kheradpour *et al.* [2013], E123/K562	FUN-LDA	0.645
	GenoSkyline	0.620
	ChromHMM	0.634
	Segway	0.585
	IDEAS	0.615
	DNase	0.656
	DNase-narrow	0.524
	DNase-gapped	0.565
	cepip_cell	0.606
	cepip_combined	0.625
dsQTLs in Degner and others [2012], E116	FUN-LDA	0.750
	GenoSkyline	0.740
	ChromHMM	0.639
	Segway	0.580
	IDEAS	0.677
	DNase	0.823
	DNase-narrow	0.665
	DNase-gapped	0.662
	cepip_cell	0.741
	cepip_combined	0.760
	deltaSVM	0.751
dsQTLs & eQTLs in Degner and others [2012], E116	FUN-LDA	0.793
	GenoSkyline	0.756
	ChromHMM	0.721
	Segway	0.648
	IDEAS	0.700
	DNase	0.832
	DNase-narrow	0.713
	DNase-gapped	0.701
	cepip_cell	0.753
	cepip_combined	0.769
	deltaSVM	0.708

I sensitivity sites that had been used to identify dsQTLs in the original study Degner and others [2012]. After removing variants with missing functional predictions, there remain 560 dsQTLs in the positive control set. We paired each of these dsQTLs with four randomly selected controls (2,236 negative controls). In addition, Degner et al. [Degner and others, 2012] observed that a substantial fraction (16%) of dsQTLs are also associated with variation in the expression levels of nearby genes (that is, these loci are also eQTLs). Therefore, we also considered separately 102 dsQTLs that are also eQTLs, and paired them with 408 randomly selected (from the set above) negative controls. We present the results in Table 4.1. It should be noted that the vast majority of dsQTLs reside close to the target DNase I hypersensitive site, and hence methods such as DNase and deltaSVM are expected to perform well for these datasets. Despite this, FUN-LDA attains an AUROC similar to deltaSVM on the dsQTL dataset, and substantially higher for the dsQTL & eQTL dataset (0.793 for FUN-LDA and 0.708 for deltaSVM).

4.3.2 Non-tissue/cell type specific validation sets

We also use several non-tissue specific datasets to validate our method. For tissue/cell type specific functional prediction methods, we construct a score by taking the maximum of the functional scores for a position across the 127 tissues in Roadmap (this is the most severe functional score for the position). We were unable to include cepip and deltaSVM in these comparisons as for these two methods scores are not available across all 127 tissues and cell types. We compare with several popular organism-level functional prediction methods, including CADD, Eigen, DANN and LINSIGHT.

We use the following lists for validation:

- 76 manually curated, experimentally validated regulatory SNPs [Li and others, 2016],
- allelic imbalanced SNPs in chromatin accessible regions from a large number of DNase-seq assays [Maurano and others, 2015],
- refined causal SNPs in non-coding regions from different sources including HGMD, ClinVar, OregAnno, and variants from fine-mapping candidate causal SNPs for 39

immune and non-immune diseases in a recent fine-mapping study [Li and others, 2016], and

- eQTLs from eleven uniformly processed fine-mapping studies [Brown *et al.*, 2016].

4.3.2.1 Validated regulatory SNPs

We used a set of 76 manually curated experimentally validated regulatory SNPs, and a set of 156 frequency-matched background SNPs within 10 kb of the curated causal variants, as used in Li and others [2016]. The results are shown in Table 4.2. As shown, FUN-LDA achieves an excellent AUROC of 0.878, substantially outperforming the organism level functional prediction methods such as CADD (0.718), Eigen (0.806), DANN (0.711), and LINSIGHT (0.818).

4.3.2.2 Allelic imbalanced SNPs in chromatin accessibility

We considered also a dataset of allelic imbalanced SNPs in chromatin accessible regions ($9,456$ positive controls, $9,678$ negative controls) identified using a large number of DNase-seq assays [Maurano and others, 2015]. The negative controls are frequency-matched background SNPs around the nearest TSS of randomly selected genes. After removing variants with missing functional predictions, there remain $8,592$ dsQTLs and $9,610$ controls. It should be noted that the allelic imbalanced SNPs are identified using DNase-seq assays, and hence DNase-max is expected to perform well for this dataset. As shown in Table 4.2, FUN-LDA performs very well with an AUROC of 0.935, higher than other tissue-specific functional prediction methods like GenoSkyline (0.906), ChromHMM (0.863), Segway (0.793) and IDEAS (0.794), and substantially better than the organism level functional prediction methods such as CADD (0.692), Eigen (0.753), DANN (0.619) and LINSIGHT (0.880).

4.3.2.3 Refined causal SNPs

We used here $5,229$ refined 'causal' SNPs in non-coding regions collected from different sources including the HGMD, ClinVar, and ORegAnno databases, and fine-mapping candi-

date causal SNPs for 39 immune and non-immune diseases from a recent fine-mapping study [Li and others, 2016]. The controls consisted of 20,916 randomly selected frequency-matched non-coding SNPs. FUN-LDA performs very well with an AUROC of 0.803, outperforming almost all the other functional prediction methods, especially the organism level prediction methods (Table 4.2): CADD (0.591), Eigen (0.655), DANN (0.587) and LINSIGHT (0.775)

4.3.2.4 Fine mapped eQTLs

Finally, we used a collection of eQTLs (31,118 positive controls, 36,540 negative controls) from the uniformly processed expression quantitative trait locus (eQTL) fine-mapping data in Brown *et al.* [2016]. The eQTLs were originally identified by multi-trait Bayesian linear regression models from eleven studies on seven tissues/cell lines, and then pre-processed by Li and others [2016] to generate a dataset of 31,118 most likely functional eQTLs (our positive controls) and 36,540 frequency-matched background SNPs around nearest TSS of randomly selected genes (our negative controls). FUN-LDA performs very well with an AUROC of 0.775, same as LINSIGHT, but substantially better than CADD (0.621), Eigen (0.653) and DANN (0.573). GenoSkyline and DNase perform slightly better than FUN-LDA for this dataset with AUROCs of 0.785 and 0.778, respectively.

4.4 Applications of our method

4.4.1 eQTL enrichment

The Genotype-Tissue Expression (GTEx) project is designed to establish a comprehensive data resource on genetic variation, gene expression and other molecular phenotypes across multiple human tissues [The GTEx Consortium, 2015]. We focus here on the cis-eQTL results from the GTEx V6 release comprising RNA-seq data on 7,051 samples in 44 tissues, each with at least 70 samples. We are interested in identifying for each GTEx tissue the Roadmap tissue that is most enriched in eQTLs from that GTEx tissue relative to other Roadmap tissues, i.e., that gives highest functional scores to those eQTLs relative to other tissues. We exclude from our analysis the sex-specific GTEx tissues (ovary, vagina, uterus, testis, prostate, breast), most of which have no relevant counterpart in Roadmap. In Table

4.3 we show the top Roadmap tissue for each remaining GTEx tissue, along with the p value from the enrichment test (see Appendix C.0.1 for details about how enrichment is measured and the p value is calculated). In most cases, eQTLs from a GTEx tissue show the most enrichment in the functional component of a relevant Roadmap tissue. For example, for liver tissue in GTEx, liver is the Roadmap tissue with the highest enrichment, for pancreas tissue in GTEx, the Roadmap tissue with the highest enrichment is pancreas, for skeletal muscle tissue in GTEx, the most enriched Roadmap tissue is skeletal muscle. However, there are also a few cases where the top tissue is not necessarily the most intuitive one, as for lung and several brain tissues. Generally, the tissues with unexpected combinations tend to either have small sample sizes for eQTL discovery in GTEx (such as brain tissues) or inadequate representation in Roadmap. Most of the mismatches have relatively large p values as well ($p > 0.001$).

4.4.2 LD score regression

As an application of the use of our scores in complex trait genetics, we use the recently developed stratified linkage disequilibrium (LD) score regression framework [Finucane *et al.*, 2015] to identify the most relevant cell types and tissues for 21 complex traits for which moderate to large GWAS studies have been performed. The stratified LD score regression approach uses information from all single nucleotide polymorphisms (SNPs) and explicitly models LD to estimate the contribution to heritability of different functional classes of variants. We modify this method to weight SNPs by their tissue specific functional score (e.g. FUN-LDA), and in this way we assess the contribution to heritability of predicted functional SNPs in a particular Roadmap cell type or tissue (see Appendix C.1 for more details).

In Table 4.4 we show the top Roadmap cell type/tissue (the one with the smallest p value from testing whether predicted functional variants in a tissue contribute significantly to SNP heritability) for each of the 21 complex traits using FUN-LDA to predict functional variants in specific cell types and tissues. For most disorders, the top tissue has previously been implicated in their pathogenesis. For example, the top tissues for body mass index (BMI) are brain tissues, consistent with recent findings indicating that BMI-associated loci

are enriched for expression in the brain and central nervous system [Locke *et al.*, 2015]. Similarly, brain represents the top tissue for most neuropsychiatric disorders, education levels, and smoking. Blood-derived and immune cells represent the top tissue for virtually all of the autoimmune conditions available for analysis. For example, GWAS findings for ulcerative colitis map specifically to the regulatory elements in Th17 cells, whereas lymphoblastoid cell lines represent the top cell type for rheumatoid arthritis. Another interesting finding involves primary hematopoietic stem cells for Alzheimer's disease, consistent with emerging data on the involvement of bone marrow-derived immune cells in the pathogenesis of neurodegeneration [Gjoneska *et al.*, 2015].

4.5 Discussion

We have introduced here a new unsupervised approach FUN-LDA for the functional prediction of genetic variation in specific cell types and tissues using histone modification and DNase data from the ENCODE and Roadmap Epigenomics projects, and have provided comparisons with commonly used functional annotation methods, both at the tissue/cell type specific and organism level. FUN-LDA is based on a mixture model that focuses on identifying the regions in the genome whose disruption is most likely to interfere with function in a particular cell type or tissue. Such context specific functional prediction of genetic variation is essential for understanding the function of noncoding variation across cell types and tissues, and for the interpretation of genetic variants uncovered in GWAS and sequencing studies. While existing segmentation approaches can be used to derive a numeric functional score as well, we have shown that they tend to be less accurate at predicting functional effects. Relative to other recently developed functional scores, such as GenoSkyline, FUN-LDA can have substantially better prediction accuracy, can use annotation data on the original scale (e.g. quantitative or binary), and makes explicit which classes are considered functionally active, namely active promoters and active enhancers, providing an attractive tool for functional scoring of variants.

In terms of prediction accuracy, we have shown that overall FUN-LDA outperforms existing methods over a variety of test datasets, sometimes substantially. In particular, we

show that compared with popular organism level functional scores such as CADD, Eigen, DANN, and LINSIGHT, FUN-LDA has substantially better accuracy. We have also shown that raw DNase can have higher predictive power than FUN-LDA and other tissue/cell type specific functional prediction methods, although the difference between FUN-LDA and DNase is minor in most comparisons, and smaller than the difference between FUN-LDA and other integrative methods (except for the DNase based datasets, such as dsQTLs and allelic imbalanced SNPs in chromatin accessibility, where DNase has an inherent advantage). This observation is concordant with a recent study showing that within open chromatin regions transcription factor binding is strongly correlated with the quantitative level of chromatin accessibility (as measured by DNase-seq) [Grossman and others, 2017]. Therefore the proposed FUN-LDA method, by being able to integrate annotation data with arbitrary distributions, has clear advantages over other mixture-based methods like GenoSkyline and ChromHMM that make use of binary peak calls. However not being a probabilistic score is a significant deficiency of DNase (e.g. enrichment analyses as shown here for eQTL, and LD score regression analyses are more difficult to implement/interpret) and in practice, in the vast majority of cases, researchers use binary DNase peak calls (DNase-narrow and DNase-gapped) rather than the quantitative DNase scores; as we show, our method FUN-LDA significantly outperforms DNase peaks on the metrics we considered.

These cell type and tissue specific functional scores have numerous applications. As shown before in Finucane *et al.* [2015], and as illustrated here as well, they can be used to infer the most relevant cell types and tissues for a trait of interest, and can help focus the search for causal variants in complex traits by restricting the set of candidate variants to only those that are predicted to be functional in tissues relevant for the trait under consideration. Beyond the application shown here, such functional predictions have numerous other applications. They can naturally be used in gene discovery studies to potentially improve power in sequence-based association tests such as SKAT and burden [Lee *et al.*, 2012; He *et al.*, 2017], and in fine-mapping studies [Ionita-Laza *et al.*, 2014; Kichaev *et al.*, 2014]. They can also be used in identifying regulatory regions that are depleted in functional variation in a specific tissue, similar to recent efforts to identify coding regions that are depleted in functional (e.g. missense, nonsense, and splice acceptor/donor variants) variation [Petro-

vski *et al.*, 2013]. Other applications include improving power of trans-eQTL studies, by using the cell type and tissue specific functional predictions as prior information. Similarly, gene-gene and gene-environment interaction studies can benefit from an analysis focused on variants predicted to be functional in a cell type or tissue relevant to the trait under study.

Choosing the number of functional classes in the LDA model is not an easy task, partly because the number of functional classes is not well defined. We have focused here on a model with nine functional classes based on biological knowledge. There is some subjectivity in any method that seeks to partition the genome into functional classes, both in terms of the number of such classes and their interpretation. Further experiments that produce catalogs of specific types of elements with validated tissue-specific functions would aid in determining the number of states that a genomic annotation model should have, and the interpretation of those states, leading to potential improvements in the accuracy of such functional predictors. Such tissue-specific experimental data would also allow the use of supervised methods which could lead to improved tissue-specific functional scores.

Unlike our method, most of the existing segmentation methods smooth the genomic signal spatially. While they thereby use information from neighboring regions in making predictions for a particular variant, they may be less able to predict functionality of narrow regions with different histone modification profiles from neighboring regions. Another difference between our method and methods that use binary peak calls is that our method can incorporate the quantitative level of the functional annotations, which can be important; for example in the case of DNase it has been recently shown that the quantitative level of chromatin accessibility is strongly correlated with transcription factor binding [Grossman and others, 2017].

We have computed FUN-LDA posterior probabilities for every position in the human genome for 127 tissue and cell types available in Roadmap. These scores are available at www.funlda.com and can also be imported into the UCSC Genome Browser.

Table 4.2: Organism level functional prediction.

Dataset	Method	AUROC
Validated regulatory SNPs	FUN-LDA-max	0.878
	GenoSkyline-max	0.846
	ChromHMM-max	0.865
	Segway-max	0.711
	IDEAS-max	0.694
	DNase-max	0.885
	DNase-narrow-max	0.828
	DNase-gapped-max	0.807
	Eigen	0.806
	CADD	0.718
	DANN	0.711
	LINSIGHT	0.818
Allelic imbalanced SNPs	FUN-LDA-max	0.935
in chromatin accessibility	GenoSkyline-max	0.906
	ChromHMM-max	0.863
	Segway-max	0.793
	IDEAS-max	0.794
	DNase-max	0.968
	DNase-narrow-max	0.869
	DNase-gapped-max	0.849
	Eigen	0.753
	CADD	0.692
	DANN	0.619
	LINSIGHT	0.880
Refined causal SNPs	FUN-LDA-max	0.803
	GenoSkyline-max	0.811
	ChromHMM-max	0.748
	Segway-max	0.714
	IDEAS-max	0.720
	DNase-max	0.807
	DNase-narrow-max	0.680
	DNase-gapped-max	0.756
	Eigen	0.655
	CADD	0.591
	DANN	0.587
	LINSIGHT	0.775
Fine mapped eQTLs	FUN-LDA-max	0.775
	GenoSkyline-max	0.785
	ChromHMM-max	0.680
	Segway-max	0.687
	IDEAS-max	0.686
	DNase-max	0.778
	DNase-narrow-max	0.615
	DNase-gapped-max	0.707
	Eigen	0.653
	CADD	0.621
	DANN	0.573
	LINSIGHT	0.777

Table 4.3: Enrichment of eQTLs among FUN-LDA predicted functional variants in tissues and cell types in Roadmap Epigenomics. The top Roadmap tissue is given for each eQTL tissue, along with the p value from a two-sample proportion test.

Tissue	Roadmap Epigenome Name	-log10(p)
Whole Blood	Primary neutrophils from peripheral blood	189.72
Cells - Transformed fibroblasts	Muscle Satellite Cultured Cells	62.69
Cells - EBV-transformed lymphocytes	GM12878 Lymphoblastoid Cells	37.74
Liver	Liver	31.82
Muscle - Skeletal	Skeletal Muscle Male	19.42
Heart - Left Ventricle	Fetal Heart	15.83
Esophagus - Mucosa	Esophagus	12.78
Pancreas	Pancreas	10.84
Colon - Transverse	Rectal Mucosa Donor 31	10.46
Artery - Tibial	Stomach Smooth Muscle	7.74
Esophagus Muscularis	Stomach Smooth Muscle	6.74
Thyroid	Fetal Intestine Small	5.96
Skin - Sun Exposed (Lower leg)	Foreskin Keratinocyte Primary Cells skin03	5.47
Spleen	Primary B cells from peripheral blood	5.35
Artery - Aorta	Aorta	5.28
Brain - Hippocampus	Brain Cingulate Gyrus	5.10
Small Intestine - Terminal Ileum	Fetal Intestine Large	5.04
Heart - Atrial Appendage	Fetal Heart	4.90
Adipose - Subcutaneous	Adipose Nuclei	4.74
Colon - Sigmoid	Colon Smooth Muscle	4.62
Brain - Caudate (basal ganglia)	Brain Substantia Nigra	4.17
Brain - Cerebellum	Adipose Derived Mesenchymal Stem Cell Cultured Cells	4.12
Nerve - Tibial	Brain Hippocampus Middle	4.11
Adrenal Gland	Fetal Adrenal Gland	3.94
Skin - Not Sun Exposed (Suprapubic)	Foreskin Keratinocyte Primary Cells skin03	3.56
Brain - Putamen (basal ganglia)	Brain Substantia Nigra	3.36
Brain - Cerebellar Hemisphere	Brain Angular Gyrus	3.08
Stomach	Stomach Mucosa	3.02
Lung	Osteoblast Primary Cells	2.57
Brain - Cortex	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells	2.10
Adipose - Visceral (Omentum)	Primary T helper cells from peripheral blood	2.00
Pituitary	Primary T helper cells PMA-I stimulated	1.96
Brain - Nucleus accumbens (basal ganglia)	H9 Cells	1.80
Esophagus - Gastroesophageal Junction	Primary neutrophils from peripheral blood	1.64
Brain - Frontal Cortex (BA9)	NHDF-Ad Adult Dermal Fibroblast Primary Cells	1.61
Artery - Coronary	Primary B cells from peripheral blood	1.35
Brain - Hypothalamus	Osteoblast Primary Cells	1.29
Brain - Anterior cingulate cortex (BA24)	A549 EtOH 0.02pct Lung Carcinoma Cell Line	1.04

Table 4.4: Top cell type/tissue in Roadmap for 21 GWAS traits using FUN-LDA posterior probabilities. The p value from the stratified LD score regression, as well as the GWAS sample size are reported for each trait.

Trait	Roadmap Epigenome Name	-log10(p)	n_{GWAS}
Schizophrenia	Fetal Brain Female	14.69	82,315
Height	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells	12.27	133,653
Rheumatoid Arthritis	GM12878 Lymphoblastoid Cells	6.92	58,284
Crohn's Disease	Primary B cells from cord blood	6.24	20,883
Age at Menarche	H9 Derived Neuronal Progenitor Cultured Cells	6.14	132,989
Educational Attainment	Fetal Brain Female	5.83	101,069
BMI	Brain Germinal Matrix	4.79	123,865
HDL	Liver	4.72	99,900
Coronary Artery Disease	Liver	4.60	86,995
Ulcerative Colitis	Primary T helper 17 cells PMA-I stimulated	4.44	27,432
Type2 Diabetes	Pancreatic Islets	4.20	69,033
Epilepsy	Brain Anterior Caudate	4.11	34,853
Triglycerides	Liver	4.10	96,598
LDL	Liver	4.08	95,454
Alopecia Areata	Primary T cells from cord blood	3.90	7,776
Alzheimer's	Primary hematopoietic stem cells G-CSF-mobilized Male	3.78	54,162
IGAN	Primary Natural Killer cells from peripheral blood	3.28	11,946
Bipolar Disorder	Fetal Brain Female	3.19	16,731
Ever Smoked	Brain Inferior Temporal Lobe	2.67	74,035
Autism	Primary monocytes from peripheral blood	2.40	10,263
Fasting Glucose	Pancreatic Islets	1.44	58,074

Part III

Bibliography

Bibliography

D Altshuler, MJ Daly, and ES Lander. Genetic mapping in human disease. *Science*, 322:881–888, 2008.

Jiawei Bai, Bing He, Haochang Shou, Vadim Zipunnikov, Thomas A Glass, and Ciprian M Crainiceanu. Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics*, 15:102–116, 2014.

AJ Bannister and T Kouzarides. Regulation of chromatin by histone modifications. *Cell Res.*, 2011.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

J Biesinger, Y Wang, and X Xie. Discovering and mapping chromatin states using a tree hidden markov model. *BMC Bioinformatics*, Suppl 5:S4, 2013.

Christopher M Bishop. Bayesian PCA. *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

S Brage, N Brage, U Ekelund, J Luan, P W Franks, K Froberg, and N J Wareham. Effect of combined movement and heart rate monitor placement on physical activity estimates during treadmill locomotion and free-living. *European Journal of Applied Physiology*, 96:517–524, 2006.

AA Brown, A Vinuela, O Delaneau, et al. Predicting causal variants affecting expression using whole-genome sequence and rna-seq from multiple human tissues. *http://www.biorxiv.org/content/biorxiv/early/2016/11/21/088872.full.pdf*, 2016.

R H Byrd, P Lu, J Nocedal, and C Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1995.

J-M Chiou, H-G Müller, and J-L Wang. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):405–423, 2003.

C Crainiceanu, P Reiss, J Goldsmith, L Huang, L Huo, and F Scheipl. *refund: Regression with Functional Data*, 2012. R package version 0.1-6.

JF Degner et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482:390–394, 2012.

C-Z Di, C M Crainiceanu, B S Caffo, and N M Punjabi. Multilevel functional principal component analysis. *Annals of Applied Statistics*, 4:458–488, 2009.

D L Donoho and V C Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems*, 16:1141–1148, 2003.

ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 2012.

J Ernst and M Kellis. Chromhmm:automating chromatin-state discovery and characterization. *Nature Methods*, 9:215–216, 2012.

J Ernst and M Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol*, 33:364–376, 2015.

HK Finucane, B Bulik-Sullivan, A Gusev, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*, 47:1228–1235, 2015.

N Friedman and OJ Rando. Epigenomics and the structure of the living genome. *Genome Res*, 25:1482–1490, 2015.

Y Fu, Z Liu, S Lu, J Bedford, X Mu, K Yip, E Khurana, and M Gerstein. Funseq2:a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15:480, 2014.

A Gelman and D B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472, 1992.

E Gjoneska, AR Pfenning, H Mathys, G Quon, A Kundaje, LH Tsai, and M Kellis. Conserved epigenomic signals in mice and humans reveal immune basis of alzheimer's disease. *Nature*, 518:365–369, 2015.

J Goldsmith and T Kitago. Assessing systematic effects of stroke on motor control using hierarchical function-on-scalar regression. *Journal of the Royal Statistical Society: Series C*, 65:215–236, 2016.

J Goldsmith, M P Wand, and C M Crainiceanu. Functional regression via variational Bayes. *Electronic Journal of Statistics*, 5:572–602, 2011.

J Goldsmith, S Greven, and C M Crainiceanu. Corrected confidence bands for functional data using principal components. *Biometrics*, 69:41–51, 2013.

J Goldsmith, V Zipunnikov, and J Schrack. Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353, 2015.

SR Grossman et al. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc Natl Acad Sci USA*, 114:E1291–E1300, 2017.

K Gu, D Pati, and D B Dunson. Bayesian hierarchical modeling of simply connected 2d shapes. *arXiv preprint arXiv:1201.1658*, 2012.

W Guo. Functional mixed effects models. *Biometrics*, 58:121–128, 2002.

P Hall, H-G Müller, and F Yao. Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B*, 70:703–723, 2008.

Z He, B Xu, S Lee, and I Ionita-Laza. Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *Am J Hum Genet*, 101:340–352, 2017.

M D Hoffman and A Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *arXiv preprint arXiv:1111.4246*, 2011.

MM Hoffman, OJ Buske, J Wang, Z Weng, J Bilmes, and WS Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*, 9:473–476, 2012.

VS Huang, SL Ryan, L Kane, S Huang, J Berard, T Kitago, P Mazzoni, and JW Krakauer. 3d robotic training in chronic stroke improves motor control but not motor function. *Society for Neuroscience. October 2012. New Orleans, USA*, 2012.

H Huang, Y Li, and Y Guan. Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, 83:210–223, 2014.

YF Huang, B Gulko, and A Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet*, 49:618–624, 2017.

I Ionita-Laza, M Capanu, S De Rubeis, K McCallum, and JD Buxbaum. Identification of rare causal variants in sequence-based studies: methods and applications to vps13b, a gene involved in cohen syndrome and autism. *PLoS Genet*, 10:e1004729, 2014.

I Ionita-Laza, K McCallum, B Xu, and JD Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*, 48:214–220, 2016.

G M James, T J Hastie, and C A Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.

C-R Jiang and J-L Wang. Covariate adjusted functional principal components analysis for longitudinal data. *The Annals of Statistics*, 38:1194–1226, 2010.

M I Jordan, Z Ghahramani, T S Jaakkola, and L K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

M I Jordan. Graphical models. *Statistical Science*, 19:140–155, 2004.

M Kellis, B Wold, MP Synder, et al. Defining functional dna elements in the human genome. *Proc Natl Acad Sci USA*, 111:6131–6138, 2014.

P Kheradpour, J Ernst, A Melnikov, P Rogov, L Wang, X Zhang, J Alston, TS Mikkelsen, and M Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*, 23:800–811, 2013.

E Khurana, Y Fu, V Colonna, et al. Integrative annotation of variants from 1092 humans:application to cancer genomics. *Science*, 342:1235587, 2013.

E Khurana, Y Fu, D Chakravarty, F Demichelis, MA Rubin, and Gerstein M. Role of non-coding sequence variants in cancer. *Nat Rev Genet*, 17:93–108, 2016.

G Kichaev, WY Yang, S Lindstrom, F Hormozdiari, E Eskin, AL Price, P Kraft, and B Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10:e1004722, 2014.

M Kircher, DM Witten, P Jain, BJ O'Roak, GM Cooper, and J Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46:310–315, 2014.

Tomoko Kitago, Jeff Goldsmith, Michelle Harran, Leslie Kane, Jessica Berard, Sylvia Huang, Sophia L. Ryan, Pietro Mazzoni, John W. Krakauer, and Vincent S. Huang. Robotic therapy for chronic stroke: general recovery of impairment or improved task-specific skill? *Journal of Neurophysiology*, 114(3):1885–1894, 2015.

J W Krakauer. Motor learning: its relevance to stroke recovery and neurorehabilitation. *Current Opinion in Neurology*, 19:84–90, 2006.

S Kurtek, A Srivastava, E Klassen, and Z Ding. Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association*, 107:1152–1165, 2012.

D D Lee and H S Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

S Lee, MC Wu, and X Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13:762–775, 2012.

D Lee, DU Gorkin, M Baker, et al. A method to predict the impact of regulatory variants from dna sequence. *Nat Genet*, 47:955–961, 2015.

MJ Li et al. Predicting regulatory variants with composite statistic. *Bioinformatics*, 32:2729–2736, 2016.

MJ Li, M Li, Z Liu, et al. cepip:context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol*, 18:52, 2017.

K Lindblad-Toh, M Garber, O Zuk, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478:476–482, 2011.

AE Locke, B Kahali, SI Berndt, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206, 2015.

Q Lu, RL Powles, Q Wang, BJ He, and Zhao H. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS Genet*, 12:e1005947, 2016.

A Mammana and HR Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol*, 16:151, 2015.

MT Maurano et al. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet*, 47:1393–1401, 2015.

M W McLean, F Scheipl, G Hooker, S Greven, and D Ruppert. Bayesian functional generalized additive models for sparsely observed covariates. *Arxiv*, 2013.

J S Morris and R J Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B*, 68:179–199, 2006.

R Neal. MCMC Using Hamiltonian Dynamics. *Handbook of Markov Chain Monte Carlo, Chapter 5*, pages 113–162, 2011.

D J Nott, M-N Tran, and C Leng. Variational approximation for heteroscedastic linear models and matching pursuit algorithms. *Statistics and Computing*, 22(2):497–512, 2012.

J Ormerod and M P Wand. Gaussian variational approximation inference for generalized linear mixed models. *The American Statistician*, 21:2–17, 2012.

J Peng and D Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18:995–1015, 2009.

S Petrovski, Q Wang, EL Heinzen, AS Allen, and Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, 9:e1003709, 2013.

D Quang, Y Chen, and X Xie. Dann:a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, 31:761–763, 2015.

J O Ramsay and B W Silverman. *Functional Data Analysis*. New York: Springer, 2005.

S Ramsey, TA Kinjnenburg, KA Kennedy, et al. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 26:2071–2075, 2010.

Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518:317–330, 2015.

F Scheipl, A-M Staicu, and S Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24:477–501, 2015.

J-P Scholz and G Schöner. The uncontrolled manifold concept: identifying control variables for a functional task. *Experimental Brain Research*, 126:289–306, 1999.

J A Schrack, V Zipunnikov, J Goldsmith, J Bai, E M Simonshick, C M Crainiceanu, and L Ferrucci. Assessing the "physical cliff": Detailed quantification of aging and physical activity. *Journal of Gerontology: Medical Sciences*, 2014.

L Shmuelof, J W Krakauer, and P Mazzoni. How is a motor skill learned? change and invariance at the levels of task success and trajectory control. *Journal of Neurophysiology*, 108(2):578–594, 2012.

BW Silverman. Density estimation for statistics and data analysis. 1986.

J Song and KC Chen. Spectacle:fast chromatin state annotation using spectral learning. *Genome Biol*, 16:33, 2015.

A Sotiras, S M Resnick, and C Davatzikos. Finding imaging patterns of structural covariance via non-negative matrix factorization. *NeuroImage*, 108:1–16, 2015.

Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 1.3*, 2013.

H Tanaka, T J Sejnowski, and J W Krakauer. Adaptation to visuomotor rotation through interaction between posterior parietal and motor cortical areas. *Journal of Neurophysiology*, 102:2921–2932, 2009.

R Tewhey et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, 165:1519–1529, 2016.

The GTEx Consortium. Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348:648–660, 2015.

M E Tipping and C Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61:611–622, 1999.

D M Titterington. Bayesian methods for neural networks and related models. *Statistical Science*, 19:128–139, 2004.

M Udell, C Horn, R Zadeh, and S Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

A van der Linde. Variational Bayesian Functional PCA. *Computational Statistics and Data Analysis*, 53:517–533, 2008.

A van der Linde. A Bayesian latent variable approach to functional principal components analysis with binary and count. *Advances in Statistical Analysis*, 93:307–333, 2009.

Václav Šmídl and Anthony Quinn. On bayesian principal component analysis. *Computational Statistics & Data Analysis*, 51:4101–4123, 2007.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B*, 73:3–36, 2011.

F. Yao, H.G. Müller, and J.L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

L Yarrow, P Brown, and J-W Krakauer. Inside the brain of an elite athlete: the neural processes that support high achievement in sports. *Nature Reviews Neuroscience*, 10:585–596, 2009.

B Zacher, M Michel, B Schwalb, P Cramer, A Tresch, and J Gagneur. Accurate promoter and enhancer identification in 127 encode and roadmap epigenomics cell types and tissues by genostan. *PLoS One*, 12:e0169249, 2017.

Y Zhang and RC Hardison. Accurate and reproducible functional maps in 127 human cell types via 2d genome segmentation. *BioRxiv*, page doi:http://dx.doi.org/10.1101/118752, 2017.

Y Zhang, L An, F Yue, and RC Hardison. Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res*, 44:6721–6731, 2016.

Part IV

Appendices

Appendix A

Appendix to Modeling motor learning using heteroskedastic functional principal components analysis

A.1 Additional results from analysis of kinematic data

One scientifically interesting question about individual motion characteristics that is addressable in our modeling framework is whether subjects with high baseline motion variance to one target tend to have high baseline motion variance to other targets. Figure A.1 shows the estimated first principal component score variance random intercept parameters $g_{il1,int}$ for each subject and each target for both the left and right hands for the X coordinate of motion, ordered by the average random intercept for each subject across targets for the right hand. There are clear subject-specific patterns of variability shared across and within hands, and clearer subject-specific patterns of variability within each hand across 8 targets. The correlation of average random intercepts for each subject across the 8 targets, one for the left and one for the right hand, was 0.56, indicating a positive correlation between baseline motor skill across hands within an individual.

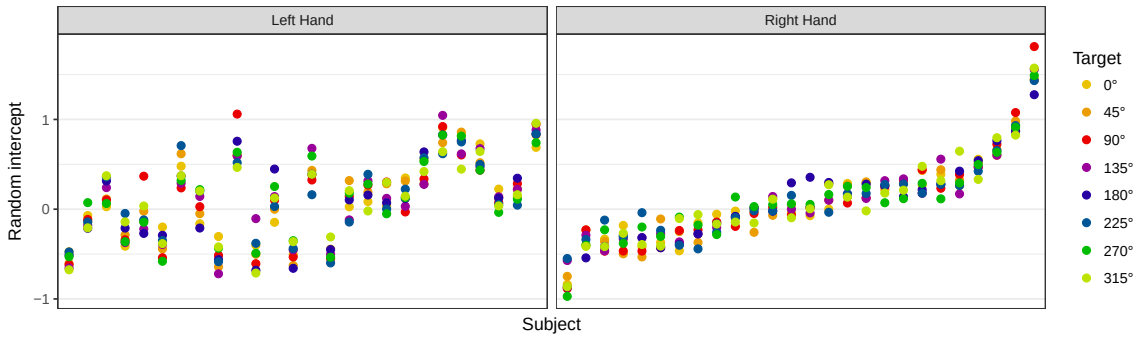

Figure A.1: Estimates of random intercepts. Each panel shows, for the left or the right hand, the estimated first principal component score variance random intercept parameters $g_{il1,int}$ in model (2.10) for each subject i and target l, for the X coordinate of motion. Targets are colored as in Figure 2.1, and subjects are ordered by their average random intercept across targets for the right hand.

Our model's point estimate of the correlation between the subject-specific cross-target score variance random intercept and the subject-specific cross-target score variance random slope is -0.80, suggesting a relationship between high baseline motion variance and faster decrease in variance with practice.

A.2 HMC and SE methods applied to kinematic data

We applied the VB, HMC and SE methods to the X coordinate of motions by the right hand to the target at 0°, and obtained very similar results. While the estimate and 95% posterior credible interval for the first FPC slope variance parameter using VB was

$$-0.020\ (-0.043, 0.003),$$

the corresponding estimate and interval for HMC was

$$-0.020\ (-0.040, -0.001)$$

and the SE confidence interval was

$$-0.023\ (-0.041, -0.005).$$

The estimates and posterior credible/confidence intervals for the first FPC intercept variance parameter were also similar:

$$3.12\ (2.81, 3.43)$$

for VB versus

$$3.18\ (2.9, 3.45)$$

for HMC and

$$3.23\ (2.97, 3.49)$$

for SE.

Estimates of random effects were also similar using the three methods, with all pairwise correlations between random intercepts and random slopes estimated using the three methods exceeding 0.85.

To generate these HMC results we ran 4 HMC chains for 2000 iterations each, and discarded the first 1000 iterations from each chain. The convergence criterion of Gelman and Rubin [1992] was less than 1.011 for each sampled variable, suggesting convergence of the chains.

A.3 Bivariate model

To fit our model to bivariate data, we make the following modifications to our model. First, $\boldsymbol{p}_{ij}$ is now a $2D \times 1$ observed functional outcome, formed by concatenating the X and Y coordinates of rotated motions. Second, our basis function matrix $\boldsymbol{\Theta}'$ is now the $2D \times 2K_\theta$ matrix $\left(\begin{smallmatrix} \boldsymbol{\Theta} & 0 \\ 0 & \boldsymbol{\Theta} \end{smallmatrix}\right)$, where $\boldsymbol{\Theta}$ is the $D \times K_\theta$ basis function matrix from model (2.5). Third, the covariance matrices in the multivariate normal distributions for $\boldsymbol{\beta}_l$, $\boldsymbol{b}_i$ and $\boldsymbol{\phi}_k$ are now the matrices (where p* represents the appropriate parameter) $\left(\begin{smallmatrix} \sigma^2_{p*,x} & 0 \\ 0 & \sigma^2_{p*,y} \end{smallmatrix}\right) \otimes \boldsymbol{P}^{-1}_{K_\theta}$, where $\otimes$ is the Kronecker product operator, $\sigma^2_{p*,x}$ and $\sigma^2_{p*,y}$ are independent with $\text{IG}\left[\alpha, \beta\right]$ priors and $\boldsymbol{P}_{K_\theta}$ is the corresponding penalty matrix from model (2.5). Finally, $\boldsymbol{\epsilon}_{ij}$ is now a $2D \times 1$ vector of independent error terms with a MVN $\left[0, \sigma^2 \boldsymbol{I}_{2D}\right]$ distribution. Since the FPCs are bi-dimensional in this model, each FPC represents a deviation from the mean motion in two dimensions, and each score represents the amount of that bi-dimensional mode of variation reflected in each motion. We assume independence of the first and last D coordinates of the functional random effects (each corresponding to a different coordinate of motion); further work could introduce correlations between them.

Figure A.2 illustrates the FPCs estimated using model (2.9) fitted to the X and Y coordinates of right hand rotated motions separately (top panels) and together using bivariate curves (bottom panels). The FPCs estimated using X and Y coordinates separately are very similar to one another. The first FPC in the bivariate model is similar to the first FPC from the model fit only to X coordinate data, and shows little variation in the Y coordinate. The second FPC in the bivariate model is similar to the first FPC from the model fit only to Y coordinate data, and shows little variation in the X coordinate. These FPCs therefore show similar patterns of variation but in different dimensions. The same pattern repeats, to a lesser extent, for the third and fourth PCs estimated using the bivariate model.

This pattern indicates that deviations from the mean motion profile in each of the dimensions represented by the X and Y coordinates are for the most part independent. The first FPC, for example, which represents a mode of variation in which motions overshoot or undershoot the target with respect to the line connecting the origin and target, is associated only with a slight systematic deviation upwards or downwards from this line. Likewise, the second FPC, which represents a mode of variation in which motions deviate upwards or

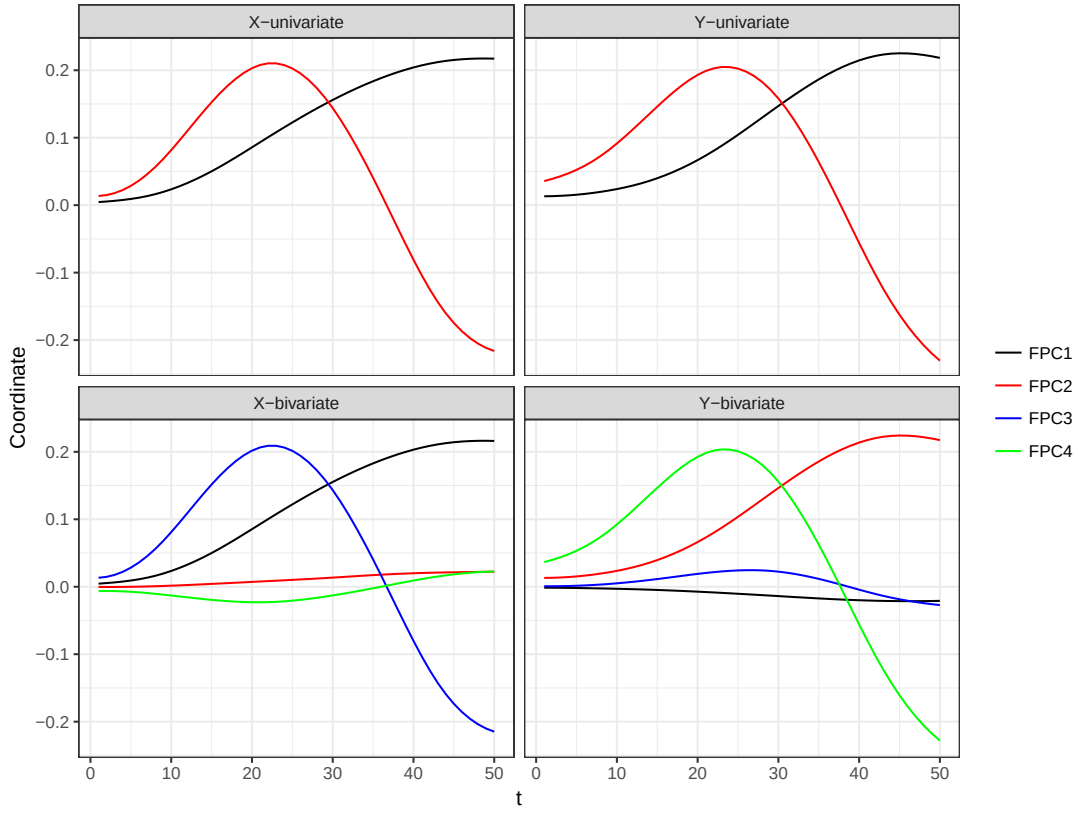

Figure A.2: FPCs from model (2.9) fit to the univariate and bivariate data. The FPCs on the left are for the X coordinates of motions, those on the right are for the Y-coordinate. The FPCs in the top row were estimated using univariate models, and the FPCs in the bottom row were estimated using bivariate models.

downwards from the line connecting the origin and the target, is associated with only a slight systematic deviation in length of motion along this line. The third and fourth FPCs represent patterns in which motions are slower than average at the beginning of the motion and then faster than average later (or vice versa). There is slightly greater involvement of both dimensions in FPCs 3 and 4.

Figure A.3 shows the change in variability of first and second bivariate FPC scores as a function of practice at the motion task. For both FPCs and all targets, score variance is estimated to decrease with motion number.

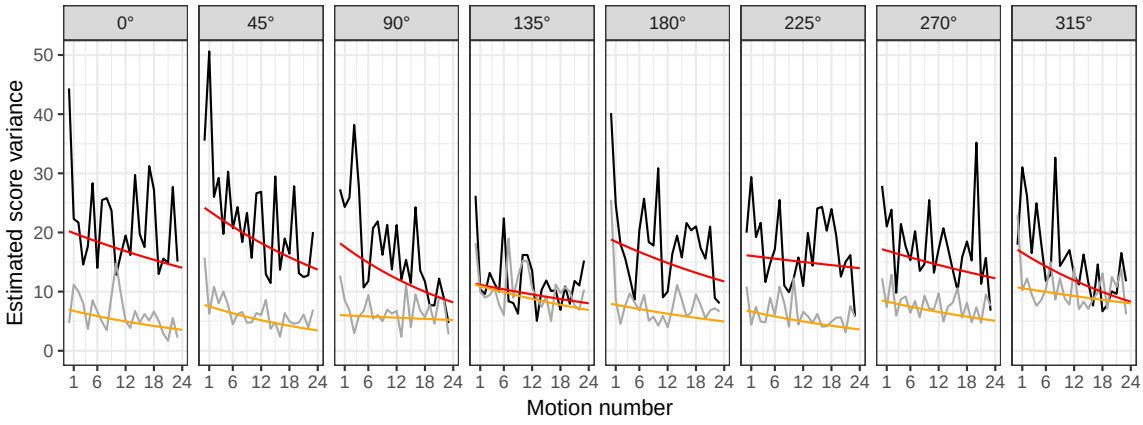

Figure A.3: Estimates of bivariate FPC score variances in the right hand for each target. Panels show the estimates of the score variance as a function of repetition number using the slope-intercept model (2.10) in red and orange (first and second FPC, respectively), and using the saturated one-parameter-per-repetition number model (2.11), in black and grey (first and second FPC, respectively).

A.4 Sensitivity Analyses

A.4.1 Hyperparameters

In our sensitivity analysis we focus on the parameters of principal interest to us in the analysis in Section 2.6, the fixed effect parameters $\gamma_{l1,slope}$, which measure how much the variability of the first FPC scores decreases with each additional motion. We found that inference for these parameters in our VB model is not sensitive to the choice of the hyper-parameters α and β in the inverse-gamma priors for the smoothing parameters $\sigma^2_{\boldsymbol{\beta}_l}$, $\sigma^2_{\boldsymbol{b}}$ and $\sigma^2_{\boldsymbol{\phi}_k}$ (we tried various combinations of values of α and β in the set $\{0.001, 0.01, 0.1, 1\}$), or to the number of spline basis functions used (we tried values in the set $\{5, 10, 15, 20\}$).

When the prior for the parameters $\gamma_{l1,int}$, which measure the baseline variance of scores for the first FPC, becomes too concentrated around zero, for example, when the variance of the mean-zero normal prior for this parameter is decreased to 1, then to compensate for the resulting severely shrunk estimates of these parameters, the estimates of $\gamma_{l1,slope}$ reverse sign. However, inference for $\gamma_{l1,slope}$ was relatively insensitive to values of the variance of this prior in the set $\{10, 100, 100\}$ (see Figures A.4 and A.5).

When using standard prior specifications for the scale matrix parameters of the inverse-Wishart priors for the random effects $\boldsymbol{g}_{ik}$ (like a diagonal identity matrix), we observed that the variance of the random effects, and credible intervals for the fixed effect parameters γ, showed dependence on the scale matrix parameters $\boldsymbol{\Psi}_k$. For this reason we use the empirical Bayes method described in Section 2.4.2.4 to set the value of these priors.

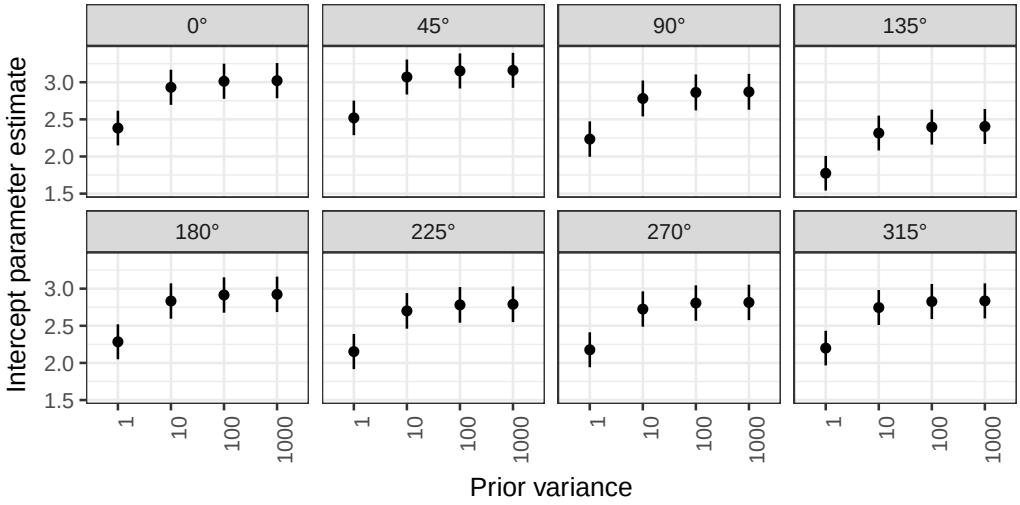

Figure A.4: Estimates and 95% credible intervals for $\gamma_{l1,int}$ as a function of the variance of its normal prior.

A.4.2 Mean Structure

We conducted various analyses to critically examine various modeling assumptions inherent in models (2.9) and (2.10). First, model (2.9) assumes that it is adequate to model the mean of the observed curves with a functional intercept for each target and random functional effects for each subject-target combination. If the mean motion to a target systematically changed as a function of repetition number, then scores at the beginning or end of the training session might be inflated, which could lead to over- or under-estimation of our parameter of principal interest, the motion number score variance slopes $\gamma_{l1,slope}$. To examine this possibility, we conducted an analysis, restricted to data for right hand motions to target 0°, in which we fit 4 separate functional random effects for each subject, for 4 groups of consecutive motions (motions 1 through 6, motions 7 through 12, et cetera). We

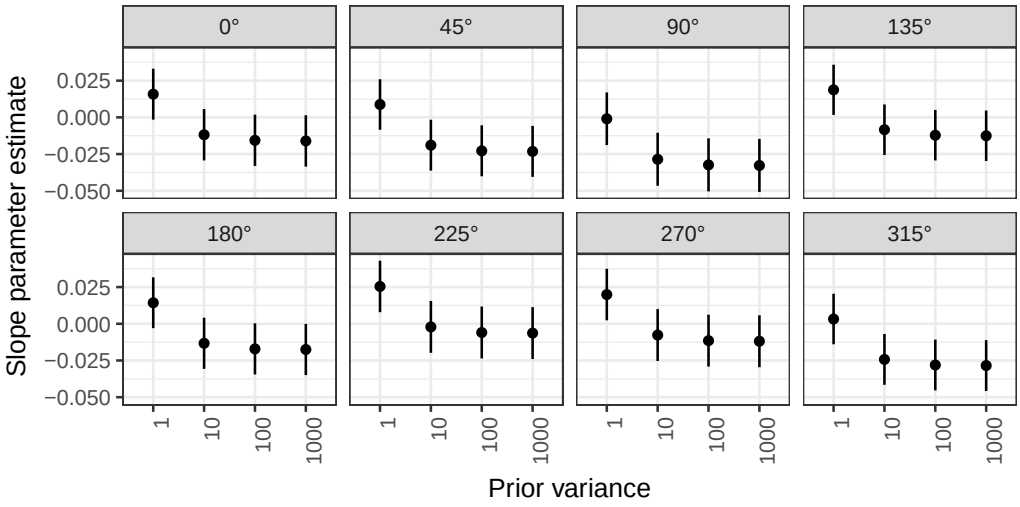

Figure A.5: Estimates and 95% credible intervals for $\gamma_{l1,slope}$ as a function of the variance of its normal prior.

found that inference for the slope parameter $\gamma_{11,slope}$ was unchanged, suggesting that model (2.9) is adequate.

Models (2.9) and (2.10) also make several simplifying independence assumptions. First, we assume independence of functional random effects for motions made by the same subject to different targets. Analysis of more complex models that modeled correlation between these functional random effects showed that although taking into account these correlations did shrink together functional random effects for the same subject, it did not change inference for our parameters of interest in the model above, the score variance repetition number slope parameters $\gamma_{l1,slope}$. Second, we assume independence of functional random effects and score variance random effects. In an ad hoc analysis to check the effects of this simplifying assumption, we included the endpoint of the estimated functional random effects as a predictor in our score variance model for data for right hand motions to target 0°. Although the 95% credible interval for this endpoint parameter did not include 0, its inclusion in the score variance model did not alter the credible interval for the repetition number slope parameter. In other contexts, for example, motions by stroke patients, correlations between functional and score variance model random effects might be stronger, and might need to be taken into account in order for inference to be correct.

A.5 Derivations

This section includes derivations of conditional distributions of all quantities in model (2.5), an overview of variational Bayes, a derivation of our variational Bayes algorithm, and additional details on the implementation of our HMC sampler. The derivations of conditional distributions are included because they are used in the derivation of our variational Bayes algorithm. Throughout this section we consider a model where each subject has one functional random effect $\boldsymbol{b}_i$. It is straightforward to extend the derivations below to the case where there are different functional random effects $\boldsymbol{b}_{im}$ for different sets of curves for each subject.

A.5.1 Derivation of conditional distributions

Let $n = \sum_{i=1}^{I} J_i$ be the total number of motions by all subjects. Let $\boldsymbol{P}$ be the $D \times n$ matrix of functional outcomes, $\boldsymbol{\beta}$ the $K_\theta \times (L+1)$ matrix of fixed effect coefficient vectors and $\boldsymbol{X}$ the corresponding $n \times (l+1)$ fixed effects design matrix, $\boldsymbol{B}$ the $K_\theta \times I$ matrix of random effect coefficient vectors and $\boldsymbol{V}$ the corresponding $n \times I$ random effects design matrix, $\boldsymbol{\Phi}$ the $K_\theta \times K$ matrix of principal component coefficient vectors and $\boldsymbol{\Xi}$ the corresponding $n \times K$ matrix of principal component scores and $\boldsymbol{E}$ the $D \times n$ error matrix of error vectors $\boldsymbol{\epsilon}_i$.

We rewrite our model using matrix notation as follows:

$$\boldsymbol{P} = \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T + \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T + \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T + \boldsymbol{E}$$

We will first derive the posterior distribution of $\boldsymbol{\beta}$ conditional on the values of the other parameters in the model. Let $\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}$ be the length $L+1$ vector of prior variances $\sigma^2_{\boldsymbol{\beta}_l}$ or, in the model with bivariate observations, the length $2L+2$ vector of prior variances $(\sigma^2_{\boldsymbol{\beta}^x_0}\sigma^2_{\boldsymbol{\beta}^y_0}, \ldots, \sigma^2_{\boldsymbol{\beta}^x_L}, \sigma^2_{\boldsymbol{\beta}^y_L})$. Let $\text{vec}\,(\boldsymbol{M})$ be the vector formed by concatenating the columns of the matrix $\boldsymbol{M}$. Then the covariance matrix of the normal prior distribution of $\text{vec}\,(\boldsymbol{\beta})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\beta}}\right) \otimes \boldsymbol{Q}^{-1}$, where $\text{diag}\,(\boldsymbol{c})$ is the matrix with the elements of $\boldsymbol{c}$ on its main diagonal and 0 elsewhere and $\otimes$ is the Kronecker product operator. The posterior distribution

of $\text{vec}(\boldsymbol{\beta})$ is then

$$\begin{aligned} p(\text{vec}(\boldsymbol{\beta})\,|\text{rest}) &\propto p(\text{vec}(\boldsymbol{P})\,|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{\Xi}, \sigma^2) p(\text{vec}(\boldsymbol{\beta})\,|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\|\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right)\|^2 + \text{vec}(\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\text{vec}(\boldsymbol{\beta})\right]\right\} \end{aligned}$$

Using the identity

$$\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B}) \tag{A.1}$$

we see that the exponent in this posterior distribution is a quadratic in $\text{vec}(\boldsymbol{\beta})$, and so the posterior distribution is multivariate normal. The inverse of the coefficient of the quadratic term is the covariance matrix of this posterior distribution:

$$\boldsymbol{\Sigma}'_{\boldsymbol{\beta}} = \left[(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}(\boldsymbol{X} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}.$$

This covariance matrix multiplied by the linear term of this exponent gives the mean of this posterior distribution:

$$\boldsymbol{\mu}'_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}'_{\boldsymbol{\beta}}(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}\left[\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right)\right].$$

The derivations of the conditional posterior distributions of $\boldsymbol{B}$ and $\boldsymbol{\Phi}$ are similar. Let $\boldsymbol{b}_i$ be the random effect for the ith subject. The covariance matrix of the normal prior distribution of $\boldsymbol{b}_i$ is $\boldsymbol{\Sigma}_{\boldsymbol{b}} = \text{diag}\left(\boldsymbol{\sigma}_{\boldsymbol{b}}^2\right) \otimes ((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})^{-1}$, where, in the model with bivariate observations, $\boldsymbol{\sigma}_{\boldsymbol{b}}^2 = (\sigma_{\boldsymbol{b}^x}^2, \sigma_{\boldsymbol{b}^y}^2)$. Let $\boldsymbol{P}_i, \boldsymbol{X}_i$ and $\boldsymbol{\Xi}_i$ be the submatrices of the matrices $\boldsymbol{P}, \boldsymbol{X}$ and $\boldsymbol{\Xi}$ corresponding to the observations for the ith subject. The posterior distribution of $\boldsymbol{b}_i$ is then

$$\begin{aligned} p(\boldsymbol{b}_i|\text{rest}) &\propto p(\text{vec}(\boldsymbol{P}_i)\,|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\Phi}, \boldsymbol{\Xi}_i, \sigma^2) p(\text{vec}(\boldsymbol{b}_i)\,|\boldsymbol{\Sigma}_{\boldsymbol{b}}) \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\|\text{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\boldsymbol{b}_i\boldsymbol{1}_{J_i}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}_i^T\right)\|^2 + \boldsymbol{b}_i^T\boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1}\boldsymbol{b}_i\right]\right\}, \end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{b}} = \left[(\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}(\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}_{\boldsymbol{b}}^{-1}\right]^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{b}_i} = \boldsymbol{\Sigma}'_{\boldsymbol{b}}(\boldsymbol{1}_{J_i} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2}\left[\text{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}_i^T\right)\right].$$

Letting $\boldsymbol{\sigma}^2_{\boldsymbol{\Phi}}$ be the length K vector of prior variances $\sigma^2_{\boldsymbol{\phi}_k}$ (or, in the model with bivariate observations, the length $2K$ vector $(\sigma^2_{\boldsymbol{\phi}^x_1}, \sigma^2_{\boldsymbol{\phi}^y_1}, \ldots, \sigma^2_{\boldsymbol{\phi}^x_K}, \sigma^2_{\boldsymbol{\phi}^y_K})$), the covariance matrix of the normal prior distribution of $\text{vec}\,(\boldsymbol{\Phi})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\Phi}} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\Phi}}\right) \otimes \boldsymbol{Q}^{-1}$. The posterior distribution of $\text{vec}\,(\boldsymbol{\Phi})$ is then

$$\begin{aligned} &p(\text{vec}\,(\boldsymbol{\Phi})\,|\text{rest}) \propto p(\text{vec}\,(\boldsymbol{P})\,|\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}, \boldsymbol{\Xi}, \sigma^2) p(\text{vec}\,(\boldsymbol{\Phi})\,|\boldsymbol{\Sigma}_{\boldsymbol{\Phi}}) \\ &\quad \propto \exp\left\{ -\frac{1}{2}\left[\frac{1}{\sigma^2} \| \text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right) \|^2 + \text{vec}\,(\boldsymbol{\Phi})^T \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Phi}} \text{vec}\,(\boldsymbol{\Phi}) \right] \right\}, \end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{\Phi}} = \left[(\boldsymbol{\Xi} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2} (\boldsymbol{\Xi} \otimes \boldsymbol{\Theta}) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Phi}} \right]^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{\Phi}} = \boldsymbol{\Sigma}'_{\boldsymbol{\Phi}} (\boldsymbol{\Xi} \otimes \boldsymbol{\Theta})^T \frac{1}{\sigma^2} \left[\text{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T - \boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T\right) \right].$$

To compute the conditional posterior distribution of $\boldsymbol{\xi}_{ij}$, the vector of scores for the jth motion for the ith subject, we let the covariance matrix of the normal prior distribution of $\boldsymbol{\xi}_{ij}$ be $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}} = \text{diag}\left(\boldsymbol{\sigma}^2_{\boldsymbol{\xi} ij}\right)$, where $\boldsymbol{\sigma}^2_{\boldsymbol{\xi}_{ij}}$ is the length K vector of prior variances for $\boldsymbol{\xi}_{ij}$. Then the posterior distribution of $\boldsymbol{\xi}_{ij}$ is

$$\begin{aligned} &p(\boldsymbol{\xi}_{ij}|\text{rest}) \\ &\quad \propto p(\boldsymbol{p}_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\Phi}, \boldsymbol{\xi}_{ij}, \sigma^2) p(\boldsymbol{\xi}_{ij}|\boldsymbol{\Sigma}_{\boldsymbol{\xi}_{ij}}) \\ &\quad \propto \exp\left(-\frac{1}{2}\left\{ \frac{1}{\sigma^2} \|\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\boldsymbol{b}_i - \boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\xi}_{ij}\|^2 + \boldsymbol{\xi}^T_{ij} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{ij}} \boldsymbol{\xi}_{ij} \right\} \right), \end{aligned}$$

that is, multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}'_{\boldsymbol{\xi}_{ij}} = \left\{ \frac{1}{\sigma^2} \boldsymbol{\Phi}^T \boldsymbol{\Theta}^T \boldsymbol{\Theta} \boldsymbol{\Phi} + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}_{ij}} \right\}^{-1}$$

and mean

$$\boldsymbol{\mu}'_{\boldsymbol{\xi}_{ij}} = \boldsymbol{\Sigma}'_{\boldsymbol{\xi}_{ij}} \boldsymbol{\Phi}^T \boldsymbol{\Theta}^T \frac{1}{\sigma^2} \left(\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\boldsymbol{b}_i \right).$$

In the model for the variance of the kth principal component scores, let $\boldsymbol{x}^*_{ijk}$ be the length $L^* + 1$ vector of fixed effect coefficients for the jth motion by the ith subject and $\boldsymbol{\gamma}_k$ the corresponding vector of fixed effects, shared across all subjects and motions, and let $\boldsymbol{z}^*_{ijk}$ be the length M^* vector of random effect coefficients for the jth motion by the ith

subject and $\boldsymbol{g}_{ik}$ the corresponding vector of random effects for the ith subject. If we let $\boldsymbol{\sigma}^2_{\gamma_k}$ be the vector of the $\sigma^2_{\gamma_{lk}}$, the prior variances of the components of γ_k, then the covariance matrix of the prior distribution of $\boldsymbol{\gamma}_k$ is $\boldsymbol{\Sigma}_{\gamma_k} = \text{diag}\left(\boldsymbol{\sigma}^2_{\gamma_k}\right)$. Let the covariance matrix of the prior distribution of $\boldsymbol{g}_{ik}$ be $\boldsymbol{\Sigma}_{g_k}$. The conditional posterior distribution of $\boldsymbol{\gamma}_k$ and the vectors $\boldsymbol{g}_{ik}, i = 1, \ldots, I$ is then

$$\begin{aligned} p(\boldsymbol{\gamma}_k, \boldsymbol{g}_{1k}, \boldsymbol{g}_{2k}, \ldots, \boldsymbol{g}_{Ik} | \text{rest}) &\propto \left(\prod_{i=1}^{I} \prod_{j=1}^{J_i} p(\xi_{ijk} | \boldsymbol{\gamma}_k, \boldsymbol{g}_{ik}) \right) p(\boldsymbol{\gamma}_k) \left(\prod_{i=1}^{I} p(\boldsymbol{g}_{ik}) \right) \\ &\propto \left(\prod_{i=1}^{I} \prod_{j=1}^{J_i} \frac{e^{-\xi^2_{ijk}/2e^{(\boldsymbol{\gamma}_k \boldsymbol{x}^*_{ijk} + \boldsymbol{g}_{ik} \boldsymbol{z}^*_{ijk})}}}{e^{(\boldsymbol{\gamma}_k \boldsymbol{x}^*_{ijk} + \boldsymbol{g}_{ik} \boldsymbol{z}^*_{ijk})/2}} \right) \exp\left[-\frac{1}{2} \left(\boldsymbol{\gamma}_k^T \boldsymbol{\Sigma}_{\gamma_k} \boldsymbol{\gamma}_k + \sum_{i=1}^{I} \boldsymbol{g}_{ik}^T \boldsymbol{\Sigma}_{g_k} \boldsymbol{g}_{ik} \right) \right], \end{aligned}$$

which has the form of the posterior of a gamma generalized linear model with log link, responses given by ξ^2_{ijk}, shape parameter equal to $1/2$ and a mean-zero multivariate normal prior on the coefficients $\boldsymbol{\gamma}_k$ and $\boldsymbol{g}_{ik}, i = 1, \ldots, I$, with covariance matrix determined by $\boldsymbol{\Sigma}_{\gamma_k}$ and $\boldsymbol{\Sigma}_{\boldsymbol{g}_k}$.

Now we derive the conditional distributions of the variance parameters, starting with $\sigma^2_{\boldsymbol{\beta}_l}$. The inverse gamma density is $p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$. Therefore the posterior distribution of $\sigma^2_{\boldsymbol{\beta}_l}$ is

$$\begin{aligned} p(\sigma^2_{\boldsymbol{\beta}_l} | \text{rest}) &\propto p(\sigma^2_{\boldsymbol{\beta}_l} | \alpha, \beta) p(\boldsymbol{\beta}_l | \sigma^2_{\boldsymbol{\beta}_l}) \\ &\propto \left(\sigma^2_{\boldsymbol{\beta}_l}\right)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{\beta}_l}}\right) \frac{1}{\left(\sigma^2_{\boldsymbol{\beta}_l}\right)^{K_\theta/2}} \exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{\beta}_l}} \boldsymbol{\beta}_l^T \boldsymbol{Q} \boldsymbol{\beta}_l\right) \\ &\propto \text{IG}\left[\alpha + \frac{K_\theta}{2}, \beta + \frac{1}{2} \boldsymbol{\beta}_l^T \boldsymbol{Q} \boldsymbol{\beta}_l\right]. \end{aligned}$$

For this variance parameter and also for the variance parameters $\sigma^2_{\boldsymbol{b}}$ and $\sigma^2_{\boldsymbol{\phi}_k}$, the conditional posterior distributions are the same in the model with bivariate observations, except that, for example, in the conditional posterior distribution of $\sigma^2_{\boldsymbol{\beta}^x_l}$, the quadratic form in the expression for the second parameter of the inverse gamma posterior distribution is computed with respect to only the first K_θ components of the vector $\boldsymbol{\beta}_l$. In the conditional distribution of $\sigma^2_{\boldsymbol{\beta}^y_l}$, the remaining components of $\boldsymbol{\beta}_l$ are used. The conditional distribution of $\sigma^2_{\boldsymbol{b}}$ is

similar:

$$\begin{aligned}
p(\sigma^2_{\boldsymbol{b}}|\text{rest}) &\propto p(\sigma^2_{\boldsymbol{b}}|\alpha,\beta)\prod_{i=1}^{I} p(\boldsymbol{b}_i|\sigma^2_{\boldsymbol{b}}) \\
&\propto \left(\sigma^2_{\boldsymbol{b}}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{b}}}\right)\frac{1}{\left(\sigma^2_{\boldsymbol{b}}\right)^{IK_\theta/2}}\exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{b}}}\sum_{i=1}^{I}\boldsymbol{b}_i^T((1-\pi)\boldsymbol{Q}+\pi\boldsymbol{I})\boldsymbol{b}_i\right) \\
&\propto \text{IG}\left[\alpha+\frac{IK_\theta}{2},\beta+\frac{1}{2}\sum_{i=1}^{I}\boldsymbol{b}_i^T((1-\pi)\boldsymbol{Q}+\pi\boldsymbol{I})\boldsymbol{b}_i\right],
\end{aligned}$$

as is the conditional distribution of $\sigma^2_{\boldsymbol{\phi}_k}$:

$$\begin{aligned}
p(\sigma^2_{\boldsymbol{\phi}_k}|\text{rest}) &\propto p(\sigma^2_{\boldsymbol{\phi}_k}|\alpha,\beta)p(\boldsymbol{\phi}_k|\sigma^2_{\boldsymbol{\phi}_k}) \\
&\propto \left(\sigma^2_{\boldsymbol{\phi}_k}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2_{\boldsymbol{\phi}_k}}\right)\frac{1}{\left(\sigma^2_{\boldsymbol{\phi}_k}\right)^{K_\theta/2}}\exp\left(-\frac{1}{2\sigma^2_{\boldsymbol{\phi}_k}}\boldsymbol{\phi}_k^T\boldsymbol{Q}\boldsymbol{\phi}_k\right) \\
&\propto \text{IG}\left[\alpha+\frac{K_\theta}{2},\beta+\frac{1}{2}\boldsymbol{\phi}_k^T\boldsymbol{Q}\boldsymbol{\phi}_k\right],
\end{aligned}$$

of σ^2:

$$\begin{aligned}
p(\sigma^2|\text{rest}) &\propto p(\sigma^2|\alpha,\beta)p(\text{vec}\,(\boldsymbol{P})\,|\boldsymbol{\beta},\boldsymbol{B},\boldsymbol{\Phi},\boldsymbol{\Xi},\sigma^2) \\
&\propto \left(\sigma^2\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)\frac{1}{\left(\sigma^2\right)^{nD/2}}\times \\
&\exp\left[-\frac{1}{2\sigma^2}\|\text{vec}\left(\boldsymbol{P}-\boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T-\boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T-\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right)\|^2\right] \\
&\propto \text{IG}\left[\alpha+\frac{nD}{2},\beta+\frac{1}{2}\|\text{vec}\left(\boldsymbol{P}-\boldsymbol{\Theta}\boldsymbol{\beta}\boldsymbol{X}^T-\boldsymbol{\Theta}\boldsymbol{B}\boldsymbol{V}^T-\boldsymbol{\Theta}\boldsymbol{\Phi}\boldsymbol{\Xi}^T\right)\|^2\right],
\end{aligned}$$

and of $\sigma^2_{g_k}$ (this is the case where there is just one scalar random effect):

$$\begin{aligned}
p(\sigma^2_{g_k}|\text{rest}) &\propto p(\sigma^2_{g_k}|\alpha,\beta)\prod_{i=1}^{I} p(g_{ik}|\sigma^2_{g_k}) \\
&\propto \left(\sigma^2_{g_k}\right)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2_{g_k}}\right)\frac{1}{\left(\sigma^2_{g_k}\right)^{I/2}}\exp\left(-\frac{1}{2\sigma^2_{g_k}}\sum_{i=1}^{I}g_{ik}^2\right) \\
&\propto \text{IG}\left[\alpha+\frac{I}{2},\beta+\frac{1}{2}\sum_{i=1}^{I}g_{ik}^2\right].
\end{aligned}$$

In our real data application, we consider a model where two random effects $g_{ik,int}$ and $g_{ik,slope}$ have a bivariate, mean-zero normal prior distribution with covariance matrix $\boldsymbol{\Sigma}_{g_k}$.

This covariance matrix has an inverse-Wishart prior distribution. The inverse-Wishart density is $p(\mathbf{\Sigma}|\mathbf{\Psi},\nu) = |\mathbf{\Sigma}|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right]\right)$, where p is the number of rows and columns of the covariance matrix $\mathbf{\Sigma}$. The conditional posterior distribution of $\mathbf{\Sigma}_{g_k}$ is therefore

$$\begin{aligned}
p(\mathbf{\Sigma}_{g_k}|\text{rest}) &\propto p(\mathbf{\Sigma}_{g_k}) \prod_{i=1}^{I} p(\boldsymbol{g}_{ik}|\mathbf{\Sigma}_g) \\
&\propto |\mathbf{\Sigma}_{g_k}|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Psi}\mathbf{\Sigma}_{g_k}^{-1}\right]\right) |\mathbf{\Sigma}_{g_k}|^{-I/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{I} \boldsymbol{g}_{ik}^T \mathbf{\Sigma}_{g_k}^{-1} \boldsymbol{g}_{ik}\right) \\
&\propto |\mathbf{\Sigma}_{g_k}|^{-\frac{\nu+p+I+1}{2}} \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{I} \text{tr}\left[\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T \mathbf{\Sigma}_{g_k}^{-1}\right] + \text{tr}\left[\mathbf{\Psi}\mathbf{\Sigma}_{g_k}^{-1}\right]\right)\right] \\
&\propto \text{IW}\left[\mathbf{\Psi} + \sum_{i=1}^{I} \boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T, \nu + I\right].
\end{aligned}$$

Straightforward extensions of these derivations apply in the case of nested random effects, as in model extension (2.6).

A.5.2 Overview of variational Bayes

Let $\boldsymbol{y}$ and $\boldsymbol{\zeta}$ represent the data and parameters, respectively, in a Bayesian model. Using variational Bayes, we approximate the posterior $p(\boldsymbol{\zeta}|\boldsymbol{y})$ using $q(\boldsymbol{\zeta})$, where q is a member of a restricted class of functions Q more easily estimated than the posterior $p(\boldsymbol{\zeta}|\boldsymbol{y})$. To find the best q in this restricted class, we choose the element $q^* \in Q$ that minimizes the Kullback-Leibler distance from $p(\boldsymbol{\zeta}|\boldsymbol{y})$. The class Q is often the class of posterior distributions satisfying some factorization property, so that $q(\boldsymbol{\zeta}) = \prod_{h=1}^{H} q_h(\zeta_h)$, with each $q_h(\zeta_h)$ a parametric density function. It can then be shown that the optimal q_h^* densities are given by

$$q_h^*(\zeta_h) \propto \exp\left[E_{-\zeta_h} \log p(\zeta_h|\text{rest})\right] \tag{A.2}$$

where $E_{-\zeta_h}$ represents the expectation with respect to the currently estimated values of all parameters except ζ_h, and "rest" represents the observed data plus all parameters other than ζ_h. This suggests the use of an iterative algorithm, setting initial values for all parameters and then updating the optimal density for each parameter ζ_h in turn, conditionally on the currently estimated values for all the other parameters.

Let $\{\sigma_s^2\}_{s\in S}$ represent the collection of all variance parameters in model (2.5). Let $\boldsymbol{\xi}_{ij}$ represent the vector of scores for the jth motion of the ith subject. The factorization we use to approximate the posterior distribution $q(\boldsymbol{\zeta})$ for model (2.5) is

$$q(\boldsymbol{\beta}_0,\ldots,\boldsymbol{\beta}_L)\left\{\prod_{i=1}^{I}\prod_{m=1}^{M} q(\boldsymbol{b}_{im})\right\} q(\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K)\left\{\prod_{i=1}^{I}\prod_{j=1}^{J_i} q(\boldsymbol{\xi}_{ij})\right\} \left\{\prod_{k=1}^{K} q(\gamma_{0k},\ldots,g_{11k},\ldots,)\right\}\left\{\prod_{s\in S} q(\sigma_s^2)\right\}$$

In the case of the model extension (2.6), each term $\boldsymbol{g}_{ik}$ would have its own factor $q(\boldsymbol{g}_{ik})$ in the factorization above.

The quality of this approximation depends on the extent to which the true posterior distribution factors as above. It is expected that the parameters in the curve mean $\mu_{ij}(t)$ and the deviation $\delta_{ij}(t)$ will be correlated, which sugg

ests that assumptions underlying the variational approximation will be violated for these components of the posterior. Nonetheless, the assumptions related to the score variance model, which is our main interest, may be sufficiently accurate to provide a reasonable approximation.

A.5.3 Derivation of variational Bayes algorithm

To find the optimal $q^*(\cdot)$ distributions for $\boldsymbol{\beta}, \boldsymbol{B}, \boldsymbol{\Phi}$ and $\boldsymbol{\Xi}$, we use the following result: if the conditional distribution of a parameter ζ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then the distribution $q^*(\zeta)$ is multivariate normal with covariance matrix $\Sigma_{q(\zeta)} = \left(E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\right]\right)^{-1}$ and mean $\mu_{q(\zeta)} = \left(E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\right]\right)^{-1} E_{-\zeta}\left[\boldsymbol{\Sigma}^{-1}\mu\right]$, where we use the notation $\mu_{q(\zeta)}$ and $\Sigma_{q(\zeta)}$, respectively, to denote the mean and variance of a parameter ζ under its optimal q^* distribution.

Throughout this section we make extensive use of the conditional distributions derived in Appendix A.5.1.

For $\mathrm{vec}\,(\boldsymbol{\beta})$, the optimal density $q^*(\mathrm{vec}\,(\beta))$ is thus multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\mathrm{vec}(\beta))} = \left[\mu_{q\left(\frac{1}{\sigma^2}\right)}((\boldsymbol{X}\otimes\boldsymbol{\Theta})^T(\boldsymbol{X}\otimes\boldsymbol{\Theta})) + \mathrm{diag}\left(\mu_{q\left(1/\sigma^2_{\beta_l}\right)}\right)\otimes\boldsymbol{Q}\right]^{-1}$$

and mean

$$\mu_{q(\mathrm{vec}(\beta))} = \boldsymbol{\Sigma}_{q(\mathrm{vec}(\beta))}(\boldsymbol{X} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\mu_{q(B)}\boldsymbol{V}^T - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q(\Xi)}^T\right)\right].$$

For $\boldsymbol{b}_i$, the optimal density $q^*(\boldsymbol{b}_i)$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(b_i)} = \left[\mu_{q\left(\frac{1}{\sigma^2}\right)}(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta}) + \mathrm{diag}\left(\mu_{q\left(1/\sigma_b^2\right)}\right) \otimes ((1-\pi)\boldsymbol{Q} + \pi\boldsymbol{I})\right]^{-1}$$

and mean

$$\mu_{q(b_i)} = \boldsymbol{\Sigma}_{q(b_i)}(\mathbf{1}_{J_i} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P}_i - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{X}_i^T - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q\left(\Xi_i^T\right)}\right)\right].$$

For $\mathrm{vec}\,(\boldsymbol{\Phi})$, the optimal density $q^*(\mathrm{vec}\,(\boldsymbol{\Phi}))$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))} = \left[\mu_{q(\boldsymbol{\Xi}^T\boldsymbol{\Xi})} \otimes (\boldsymbol{\Theta}^T\boldsymbol{\Theta}) + \mathrm{diag}\left(\mu_{q\left(1/\sigma_\Phi^2\right)}\right) \otimes \boldsymbol{Q}\right]^{-1}$$

and mean

$$\mu_{q(\mathrm{vec}(\Phi))} = \boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}(\mu_{q(\Xi)} \otimes \boldsymbol{\Theta})^T \mu_{q\left(\frac{1}{\sigma^2}\right)} \left[\mathrm{vec}\left(\boldsymbol{P} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{X}^T - \boldsymbol{\Theta}\mu_{q(B)}\boldsymbol{V}^T\right)\right].$$

For $\boldsymbol{\xi}_{ij}$, letting $\mu_{q\left(\Sigma_{\xi_{ij}}^{-1}\right)}$ represent the expectation under the current distributions of the parameters γ_{lk} and g_{imk} of the precision matrix of the $\boldsymbol{\xi}_{ij}$, the optimal density $q^*(\boldsymbol{\xi}_{ij})$ is multivariate normal with covariance matrix

$$\boldsymbol{\Sigma}_{q(\xi_{ij})} = \left\{\mu_{q\left(\frac{1}{\sigma^2}\right)}\mu_{q(\boldsymbol{\Phi}^T\boldsymbol{\Theta}^T\boldsymbol{\Theta}\boldsymbol{\Phi})} + \mu_{q\left(\Sigma_{\xi_{ij}}^{-1}\right)}\right\}^{-1}$$

and mean

$$\mu_{q(\xi_{ij})} = \boldsymbol{\Sigma}_{q(\xi_{ij})}\mu_{q(\Phi)}^T\boldsymbol{\Theta}^T\mu_{q\left(\frac{1}{\sigma^2}\right)}\left(\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\mu_{q(b_i)}\right).$$

The expectation $\mu_{q(\Phi^T\Theta^T\Theta\Phi)}$ appearing in the above expression for $\boldsymbol{\Sigma}_{q(\xi_{ij})}$ is the $K \times K$ matrix given by $\mu_{q(\Phi)}^T\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mu_{q(\Phi)} + \{M_{ij}\}$ where $M_{ij} = \mathrm{tr}\left[\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mathrm{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)\right]$ and $\mathrm{cov}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j)$ is a submatrix of $\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}$. The expectation $\mu_{q(\Xi^T\Xi)}$ appearing in the above expression for $\boldsymbol{\Sigma}_{q(\mathrm{vec}(\Phi))}$ is the $K \times K$ matrix given by $\mu_{q(\Xi)}^T\mu_{q(\Xi)} + M$, where $M = \sum_{i,j}\boldsymbol{\Sigma}_{q(\xi_{ij})}$.

Let $(\boldsymbol{\gamma}, \boldsymbol{g})_k$ represent the vector $(\boldsymbol{\gamma}_k, \boldsymbol{g}_{1k}, \boldsymbol{g}_{2k}, \ldots, \boldsymbol{g}_{Ik})$. As in Nott *et al.* [2012], we use a multivariate normal approximation to the density $q((\boldsymbol{\gamma}, \boldsymbol{g})_k)$. Using a routine from Nott *et al.* [2012], we approximate the mean $\mu_{q((\boldsymbol{\gamma},\boldsymbol{g})_k)}$ of the density $q((\boldsymbol{\gamma}, \boldsymbol{g})_k)$ with the posterior mode of the Bayesian gamma generalized linear model corresponding to the conditional

posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{g})_k$, using as responses the expectations $\mu_{q(\xi^2_{ijk})}$ in place of ξ^2_{ijk}, and we approximate the variance $\Sigma_{q((\boldsymbol{\gamma},\boldsymbol{g})_k)}$ with the negative inverse Hessian of the log posterior at the mode. Let these approximations be $\boldsymbol{\mu}_{mode}$ and $\boldsymbol{\Sigma}_{mode}$. Then, if ξ_{ijk} has the distribution $N[0, \exp\left(\boldsymbol{x}^T(\boldsymbol{\gamma}, \boldsymbol{g})_k\right)]$ for some coefficient vector $\boldsymbol{x}$, then by completing the square, we find that the expectation $\mu_{q\left(\Sigma^{-1}_{\xi_i j}\right)}$ in the expression for $\boldsymbol{\Sigma}_{q(\xi_{ij})}$ above is $\exp\left(-\boldsymbol{\mu}^T_{mode}\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_{mode}\boldsymbol{x}\right)$.

To find the optimal $q^*(\cdot)$ distributions for $\sigma^2_{\boldsymbol{\beta}_l}$, $\sigma^2_{\boldsymbol{b}}$, $\sigma^2_{\boldsymbol{\phi}_k}$ and σ^2, we use the following result: if the conditional distribution of a parameter ζ is inverse gamma with parameters α and β, then the distribution $q^*(\zeta)$ is inverse gamma with parameters $E_{-\zeta}[\alpha]$ and $E_{-\zeta}[\beta]$, and the expectation $\mu_{q(1/\zeta)}$ is $E_{-\zeta}[\alpha]/E_{-\zeta}[\beta]$.

For $\sigma^2_{\boldsymbol{\beta}_l}$, the optimal density $q^*(\sigma^2_{\boldsymbol{\beta}_l})$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\beta_l^T Q \beta_l\right)}$. For $\sigma^2_{\boldsymbol{b}}$, the optimal density $q^*(\sigma^2_{\boldsymbol{b}})$ is inverse gamma with parameters $\alpha + \frac{IK_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\sum_{i=1}^I b_i^T((1-\pi)Q+\pi I)b_i\right)}$. For $\sigma^2_{\boldsymbol{\phi}_k}$, the optimal density $q^*(\sigma^2_{\boldsymbol{\phi}_k})$ is inverse gamma with parameters $\alpha + \frac{K_\theta}{2}$ and $\beta + \frac{1}{2}\mu_{q\left(\phi_k^T Q \phi_k\right)}$. All of these expectations can be found using the optimal $q^*()$ distributions for $\boldsymbol{\beta}_l$, $\boldsymbol{b}_i$ and $\boldsymbol{\phi}_k$ and the formula for the expectation of a quadratic form.

For σ^2, let $\boldsymbol{x}_{ij}$ be the row of the matrix $\boldsymbol{X}$ corresponding to the jth motion of the ith subject. Then the optimal density $q^*(\sigma^2)$ is inverse gamma with parameters $\alpha + \frac{nD}{2}$ and

$$\beta + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left[\|\boldsymbol{p}_{ij} - \boldsymbol{\Theta}\mu_{q(\beta)}\boldsymbol{x}_{ij} - \boldsymbol{\Theta}\mu_{q(b_i)} - \boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q(\xi_{ij})}\|^2 + \boldsymbol{x}_{ij}\boldsymbol{L}\boldsymbol{x}_{ij}^T + m_i + n_{ij}\right]$$

where the matrix $\boldsymbol{L}$ is the $(l+1)\times(l+1)$ matrix whose i,j entry is the trace of $\boldsymbol{\Theta}^T\boldsymbol{\Theta}$ times the covariance between the ith and jth column of $\boldsymbol{\beta}$ under the current distribution of $\boldsymbol{\beta}$, $m_i = \mathrm{tr}\left[\boldsymbol{\Theta}^T\boldsymbol{\Theta}\boldsymbol{\Sigma}_{q(b_i)}\right]$, and

$$n_{ij} = \mu^T_{q(\xi_{ij})}\mu_{q(\Phi^T\Theta^T\Theta\Phi)}\mu_{q(\xi_{ij})} + \mathrm{tr}\left[\mu_{q(\Phi^T\Theta^T\Theta\Phi)}\boldsymbol{\Sigma}_{q(\xi_{ij})}\right] - \mu^T_{q(\xi_{ij})}\mu^T_{q(\Phi)}\boldsymbol{\Theta}^T\boldsymbol{\Theta}\mu_{q(\Phi)}\mu_{q(\xi_{ij})}.$$

The optimal $q^*(\Sigma_{g_k})$ density is given by

$$\begin{aligned} q^*(\Sigma_{g_k}) &\sim \exp[E_{-\Sigma_{g_k}}\log p(\Sigma_{g_k}|\text{rest})] \\ &\sim \exp\left[E_{-\Sigma_{g_k}}\left\{-\frac{\nu+I+p+1}{2}\log|\Sigma| - \frac{1}{2}\left(\mathrm{tr}\left[(\Psi + \sum_{i=1}^{I}\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)\Sigma^{-1}\right]\right)\right\}\right] \end{aligned}$$

Therefore the optimal density is inverse-Wishart with parameters $\nu+I$ and $\boldsymbol{\Psi}+\sum_{i=1}^{I}\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}$. The expectation $\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}$ in this expression is $\mu_{q(\boldsymbol{g}_{ik})}\mu_{q(\boldsymbol{g}_{ik})}^T+M$, where M is the covariance of $\boldsymbol{g}_{ik}$ under the posterior distribution of $(\boldsymbol{\gamma},\boldsymbol{g})_k$. The mean of this density is

$$\mu_{q(\Sigma_{gk})}=\frac{\boldsymbol{\Psi}+\sum_{i=1}^{I}\mu_{q(\boldsymbol{g}_{ik}\boldsymbol{g}_{ik}^T)}}{\nu+I-p-1}.$$

Straightforward extensions of these derivations apply in the case of nested random effects, as in model extension (2.6).

A.5.4 Details of implementation of HMC sampler

Our HMC samplers in Sections 3.4 and 2.6 fit the same models as fit by our VB model, while conditioning on VB estimates of the parameters $\boldsymbol{\beta}_l$, $\boldsymbol{b}_{im}$ and $\boldsymbol{\phi}_k$ in model (2.5), and therefore implicitly also conditioning on the associated variance parameters and on the VB estimate of π. The HMC samplers estimate all other parameters in these models: the scores ξ_{ijk}, the fixed effect variance parameters γ_{lk}, the random effect variance parameters $\boldsymbol{g}_{ik}$ (and $\boldsymbol{g}_{ilk}$, in model extension (2.6)), the random effect variance parameter covariance matrices, and the error variance σ^2. The samplers were implemented in the `STAN` Bayesian programming language [Stan Development Team, 2013]. `STAN` implements Hamiltonian Monte Carlo, an MCMC algorithm that uses the gradient of the log-posterior to avoid random walk behavior and therefore more quickly generate samples from the posterior [Neal, 2011].

We ran all HMC samplers here using 4 chains and checked for convergence using the convergence criterion of Gelman and Rubin [1992]. We ran the HMC sampler used in Section 3.4 for 800 iterations per chain, and discarded the first 400 iterations from each chain, which took about 90 minutes per chain. We ran the HMC sampler used in Appendix A.2 for 2000 iterations per chain, and discarded the first 1000 iterations from each chain.

Code implementing the `STAN` model used in Section 3.4 is included in the Supplementary Materials.

A.6 Additional simulation results

Here we present cross-sectional simulations to illustrate the effect of varying the number of curves, the number of estimated FPCs, the number of spline basis functions and the measurement error on the quality of estimation using the VB method. In this cross-sectional design, curves are generated from the model

$$P_i(t) = 0 + \sum_{k=1}^{4} \xi_{ik}\phi_k(t) + \epsilon_i(t).$$

FPCs and group and FPC-specific score variances are as in the simulations in Section 3.4.

All results are for 200 replicates per simulation scenario. We present one simulation where we fix the number of estimated FPCs at 4, the number of spline basis functions at 10, and the measurement error standard deviation at 0.25, and vary the number of curves in the set $\{20, 40, 80, 160, 320\}$. In the other simulations we fix the sample size at 80 and vary one of the other parameters.

For each simulated dataset, we use the methods described in Section 3.2 to fit the model

$$\boldsymbol{p}_i = \boldsymbol{\Theta}\boldsymbol{\beta}_0 + \sum_{k=1}^{K} \xi_{ik}\boldsymbol{\Theta}\boldsymbol{\phi}_k + \boldsymbol{\epsilon}_i \tag{A.3}$$

$$\xi_{ik} \sim \mathrm{N}\left[0, \exp\left(\sum_{m=1}^{2} \gamma_{lk} x^*_{il}\right)\right]. \tag{A.4}$$

The covariates x^*_{il} are defined like the analogous covariates in Section 3.4.

Figure A.6 shows that accuracy in estimation of FPCs and bias in estimation of variance model parameters decreases with more curves. Figure A.7 shows that when 2 or 3 FPCs are estimated instead of the 4 that actually exist, estimates of the quantities that are estimated are not negatively affected. Figure A.8 shows the result of changing the number of spline basis functions used for estimation. 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4; otherwise, because we induce smoothness in the estimated FPCs using the penalty matrix $\boldsymbol{Q}$, using richer spline bases does not negatively affect estimation accuracy. Figure A.9 shows the result of adding more noise to the simulated curves, keeping the sample size fixed. As expected, more noise results in larger errors in estimation, of both the FPCs and the score variance parameters.

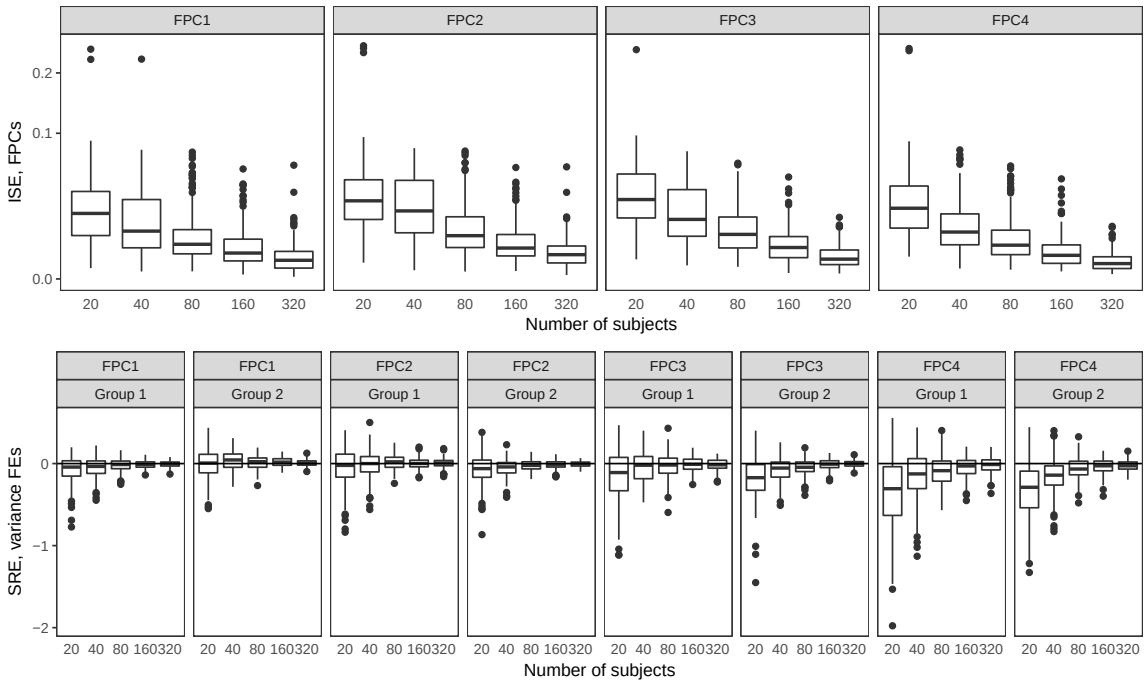

Figure A.6: Varying the number of curves. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) decreases with more curves.

Figure A.10 shows examples of estimates of FPC 2 with varying levels of integrated squared error. These estimates are from the longitudinal simulation scenario with $J_i = 4$.

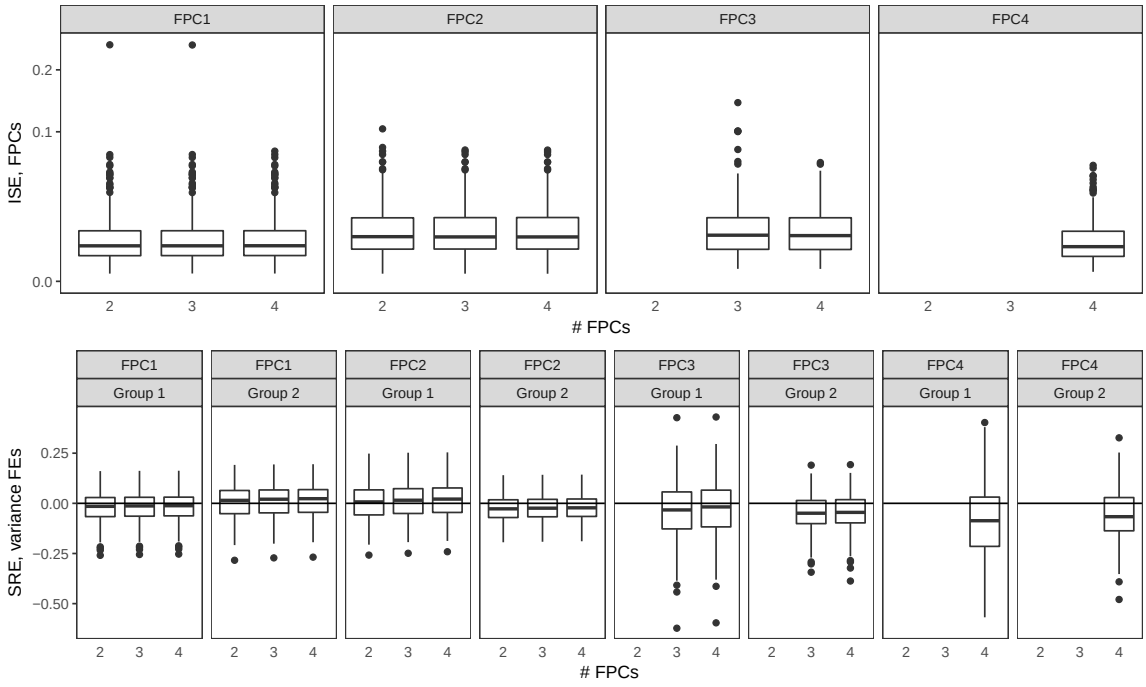

Figure A.7: Varying the number of estimated FPCs. Integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) for FPCs 1 and 2 is mostly invariant to whether additional FPCs and associated score variance parameters are also estimated.

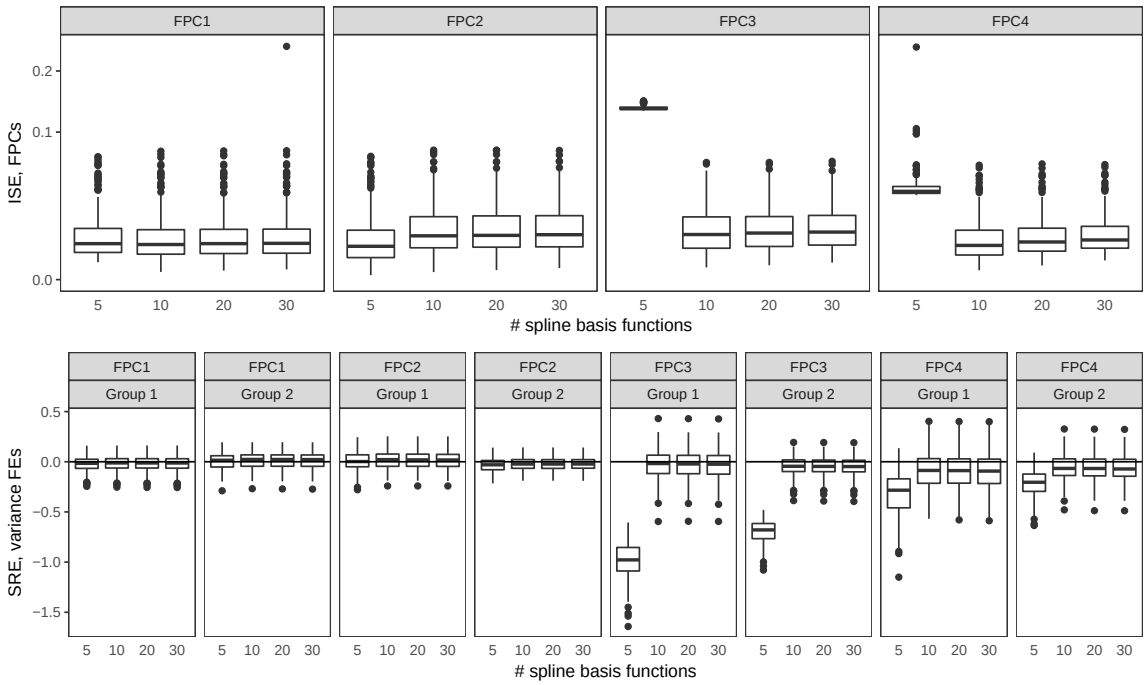

Figure A.8: Varying the number of spline basis functions. 5 spline basis functions are not sufficient to adequately capture the relatively fast variation in FPCs 3 and 4. Otherwise integrated squared errors in estimation of FPCs (first row) and signed relative error in estimation of variance parameters (second row) are mostly invariant to the number of spline basis functions used in simulation.

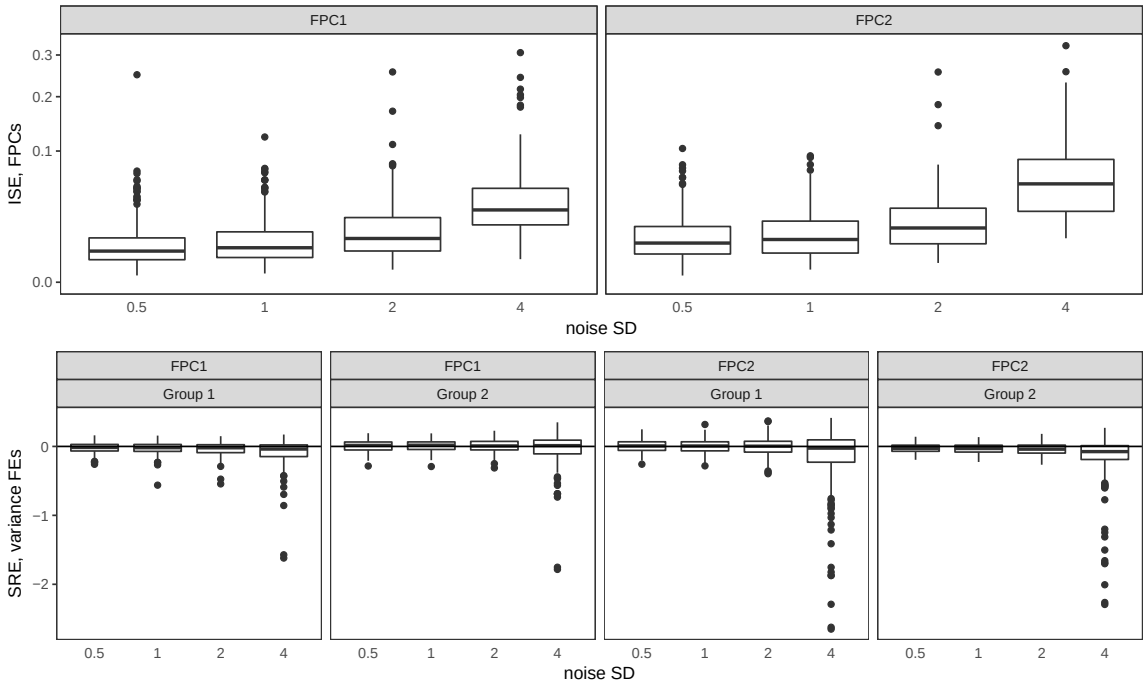

Figure A.9: Varying the measurement error. We varied the measurement error standard deviation to 0.5, 1, 2 and 4. FPC integrated squared errors (first row) and signed relative errors in estimation of the variance parameters (second row) illustrate that results are robust to a significant amount of noise, but estimation of parameters becomes poorer as the amount of noise increases. Four FPCs were simulated but only 2 were estimated.

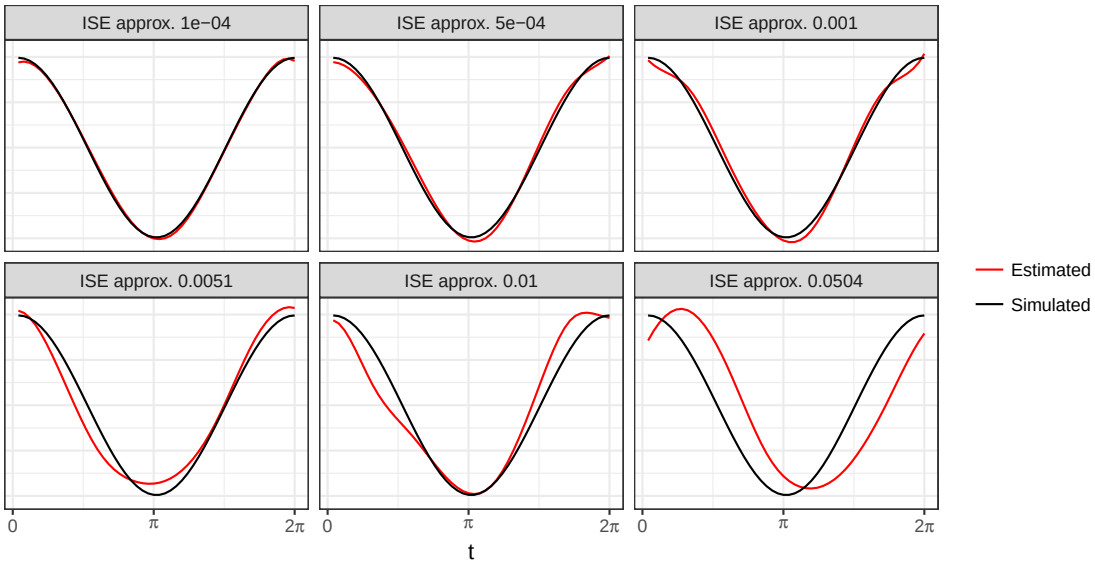

Figure A.10: Examples of estimates of FPC 2 with varying levels of integrated squared error. These estimates come from the longitudinal simulation scenario with $J_i = 4$.

Appendix B

Appendix to Non-negative matrix factorization approach to analysis of functional data

B.1 Additional figures

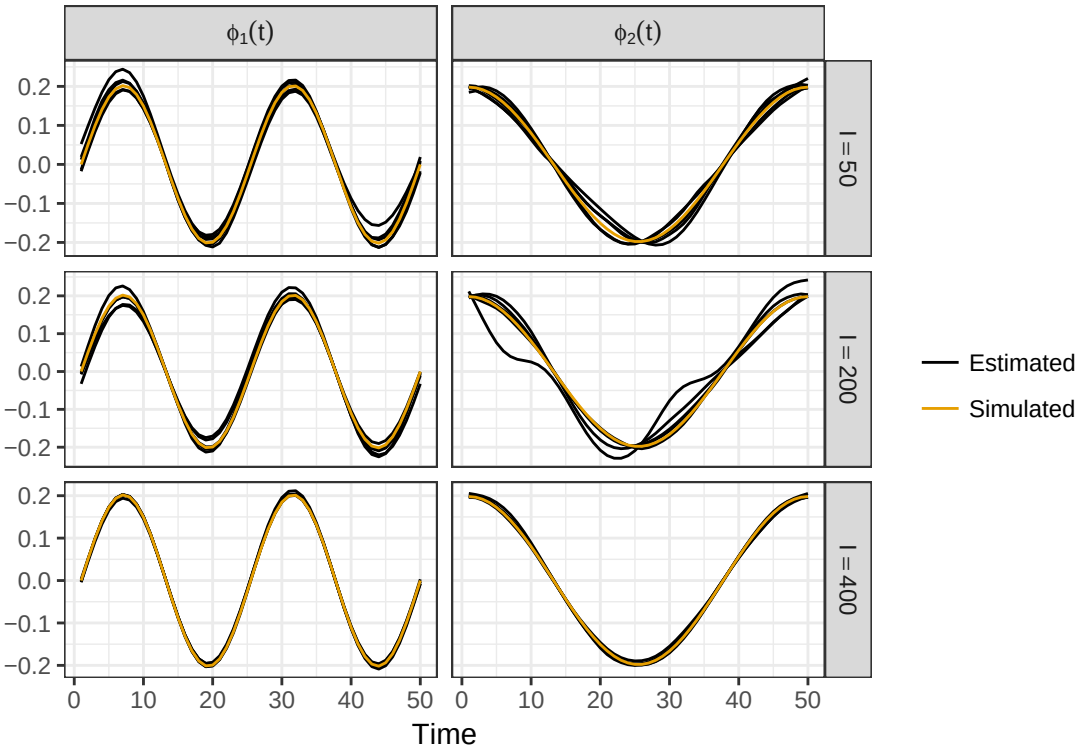

Figure B.1: Simulated FPCs and GFPCA estimates for Scenario II, for different numbers of curves per simulation replicate. Each simulation was replicated 5 times.

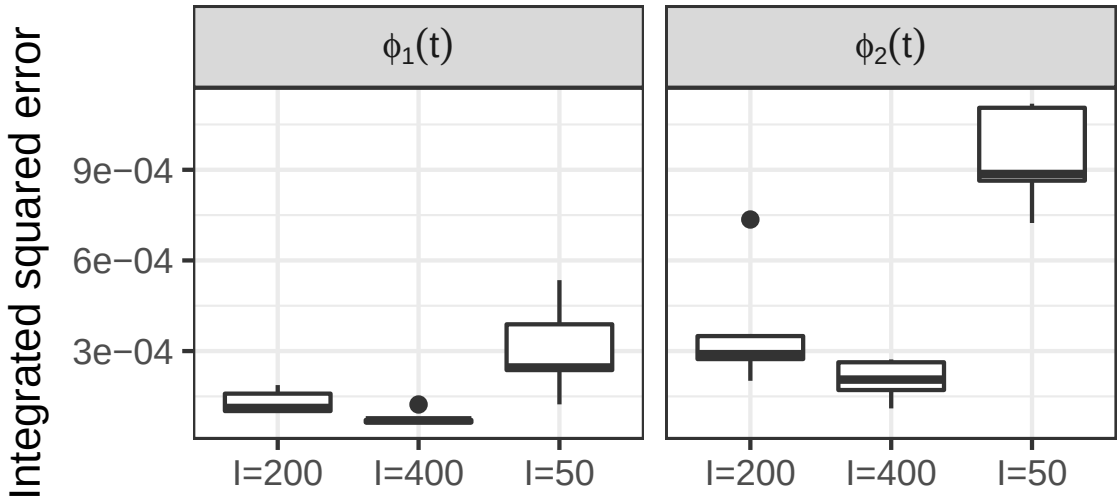

Figure B.2: Integrated squared errors of estimation of functional prototypes estimated using NARFD for $I \in \{50, 200, 400\}$ and simulation Scenario I.

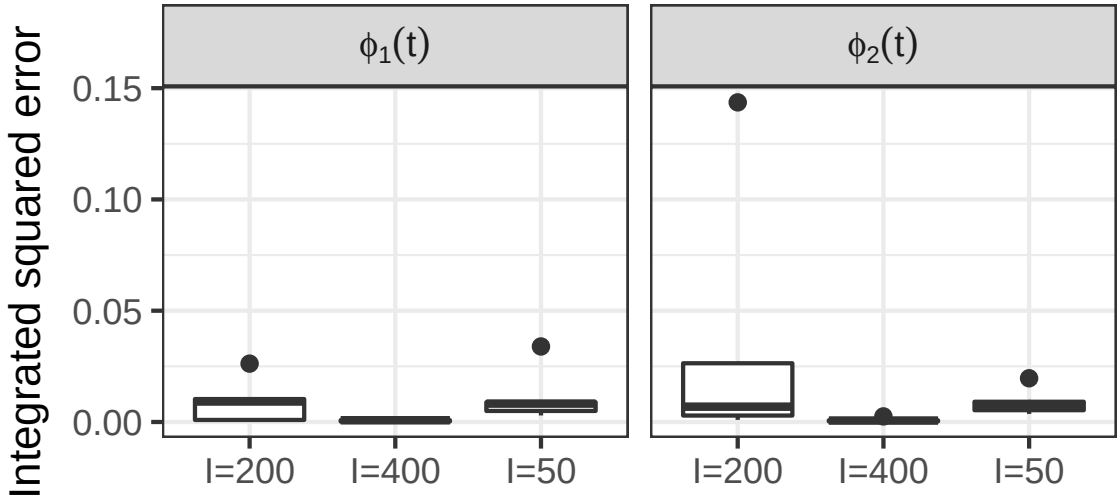

Figure B.3: Integrated squared errors of estimation of FPCs estimated using GFPCA for $I \in \{50, 200, 400\}$ and simulation Scenario II.

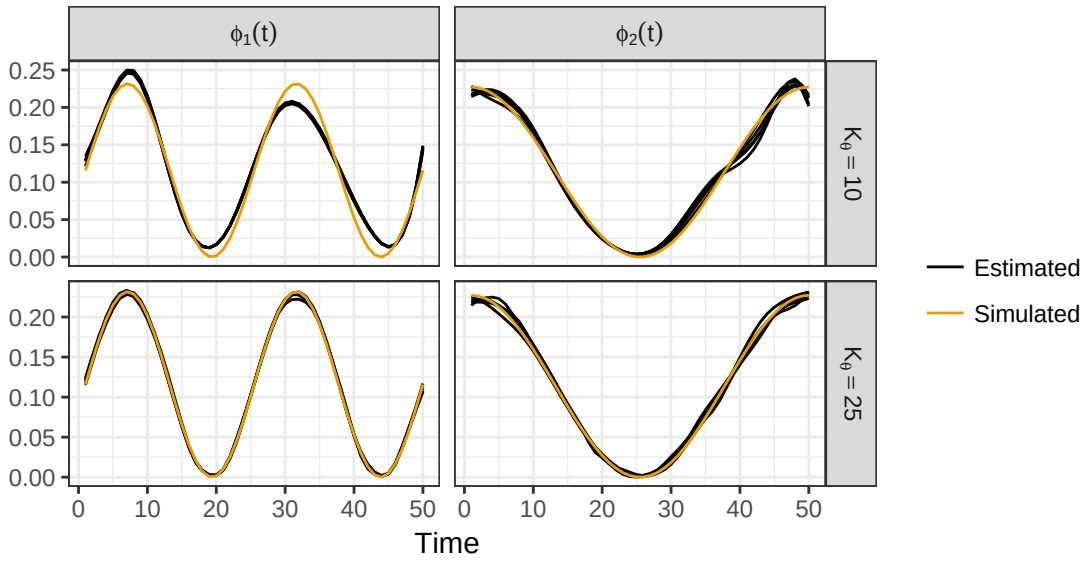

Figure B.4: Effect of changing K_θ on NARFD estimation, with $I = 50$ and simulation Scenario I.

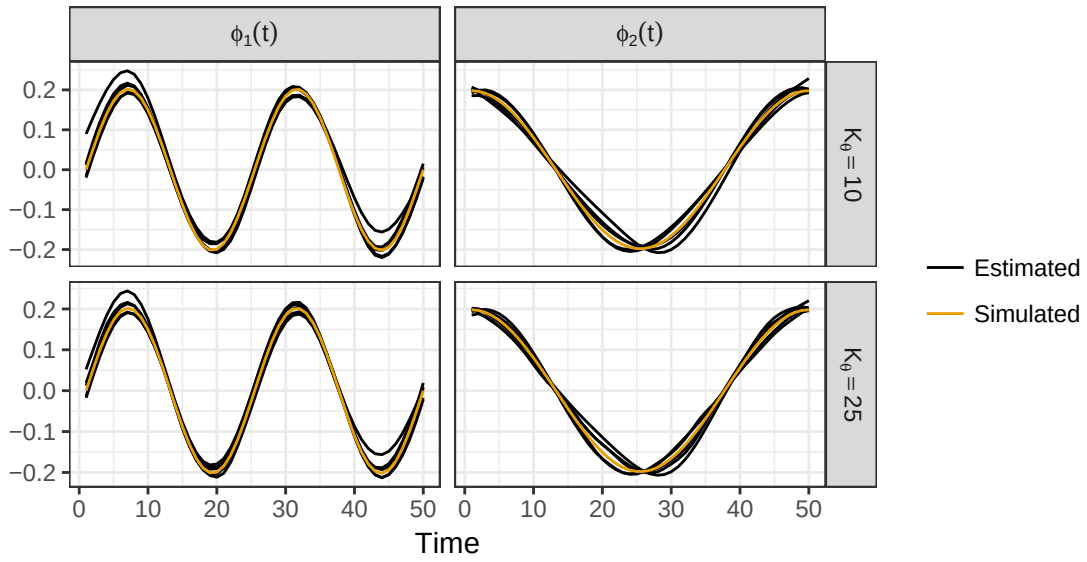

Figure B.5: Effect of changing K_θ on GFPCA estimation, with $I = 50$ and simulation Scenario II.

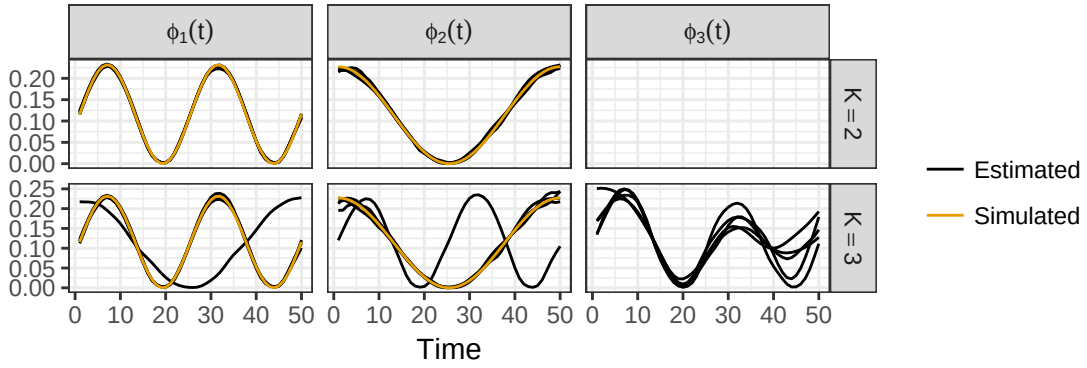

Figure B.6: Effect of estimating more functional prototypes than simulated with NARFD, with $I = 50$ and simulation Scenario I. Two functional prototypes were simulated and, in the bottom panel, three were estimated. Estimated functional prototypes are labeled based on their total contribution to the curve reconstructions. Since the contribution of the high frequency cosine to the reconstructions is now split among two prototypes, the order of the first two prototypes is sometimes switched.

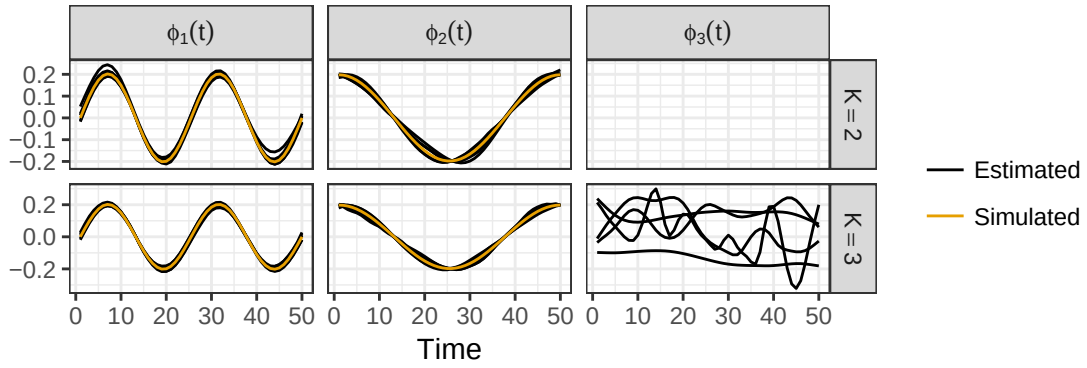

Figure B.7: Effect of estimating more FPCs than used in simulation on GFPCA estimation, with $I = 50$ and simulation Scenario II. Two FPCs were simulated and, in the bottom panel, three were estimated. Estimated FPCs are labeled based on the variance of their scores, after the FPCs have been normalized to have unit norm.

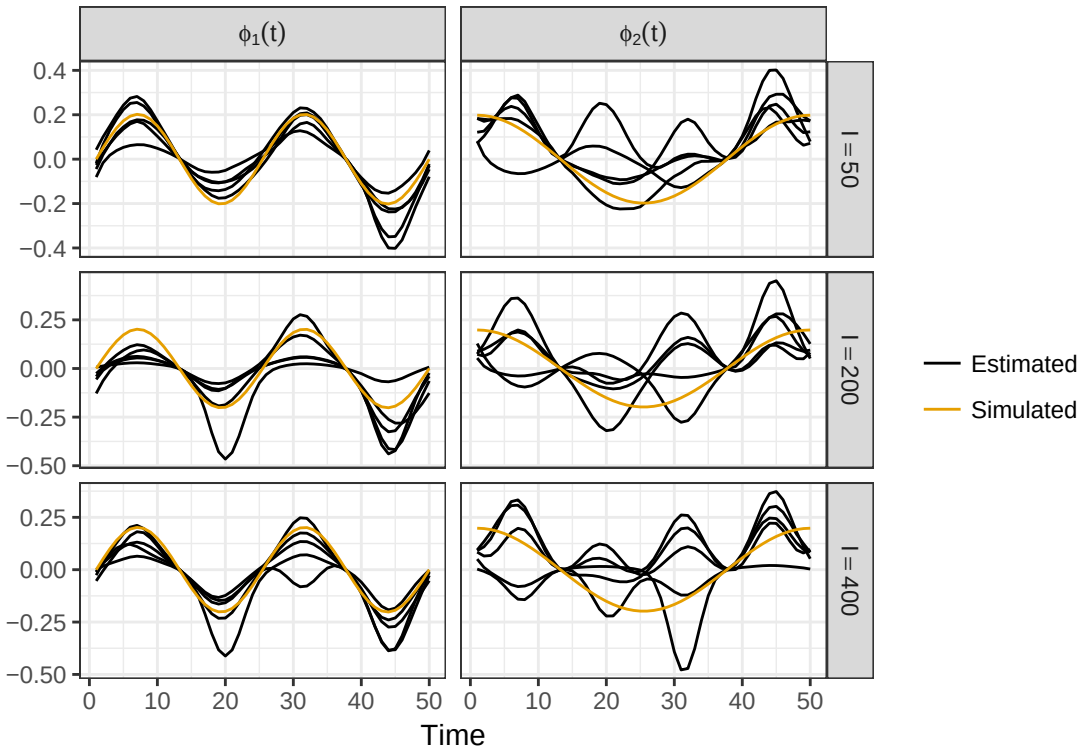

Figure B.8: Simulated FPCs and estimates using the method of Hall *et al.* [2008] for Scenario II, for different numbers of curves per simulation replicate. Each simulation was replicated 5 times. The poor performance of this method may be due to a violation of its assumption that the variation of the curves about the mean is relatively small.

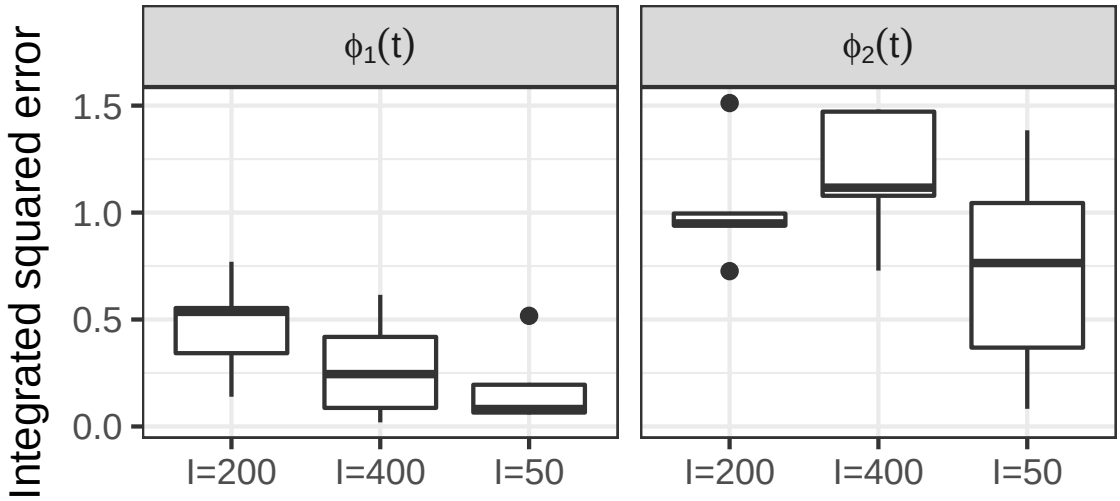

Figure B.9: Integrated squared errors of estimation of FPCs estimated using the method of Hall *et al.* [2008] for $I \in \{50, 200, 400\}$ and simulation Scenario II.

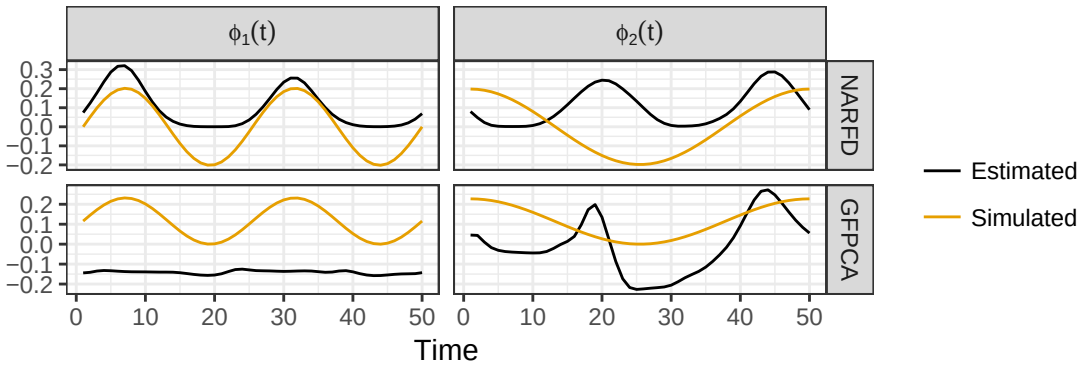

Figure B.10: The top panel shows FPCs simulated under Scenario II (the GFPCA generative model) and corresponding functional prototypes estimated with NARFD. The bottom panel shows functional prototypes simulated under Scenario I (the NARFD generative model) and corresponding FPCs estimated with GFPCA.

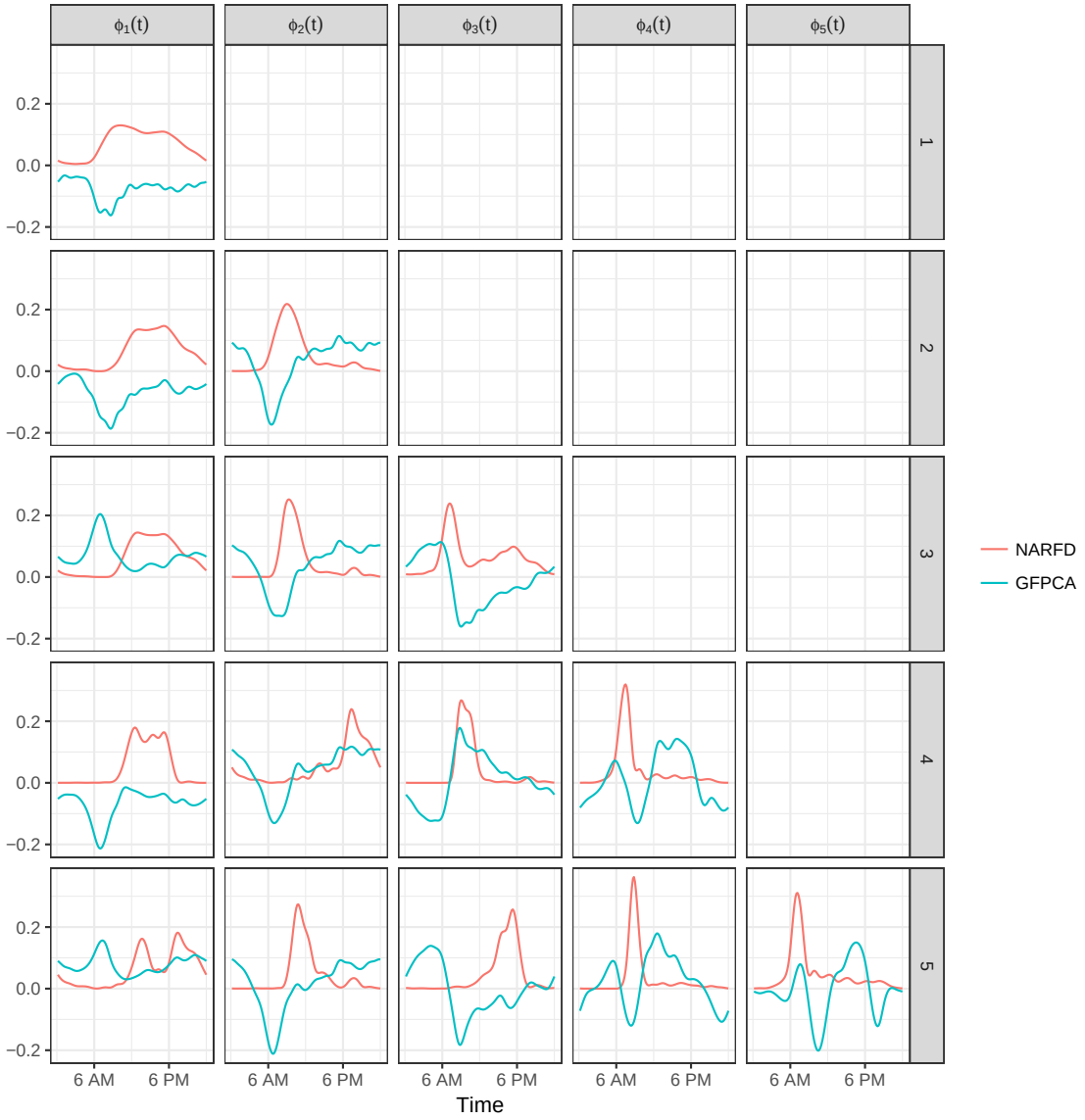

Figure B.11: Functional prototypes and FPCs estimated using BLSA data using 1 through 5 FPCs/prototypes, using NARFD and GFPCA. For both methods, the kth estimated FPC/prototype is not invariant to how many FPCs/prototypes are estimated. GFPCA FPCs are shown on the scale on which they are estimated.

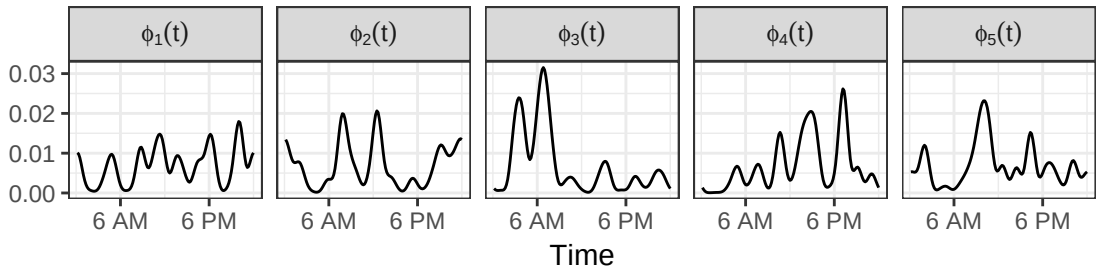

Figure B.12: Five functional prototypes estimated using BLSA data using non-negative matrix factorization, without any smoothing, using data from all 592 subjects.

Appendix C

Appendix to FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation

C.0.1 eQTL enrichment

Let $G_1, \ldots, G_{44}$ be the 44 GTEx tissues with at least 70 samples, and $R_1, \ldots, R_{127}$ be the 127 Roadmap tissues. For a given tissue in GTEx G_i we are interested in identifying the Roadmap tissue R_j with the highest enrichment in eQTLs from G_i relative to other tissues in Roadmap.

Let

$$p_{G_i|R_j} = \frac{\#\text{eQTLs in tissue } G_i \text{ in functional component of } R_j}{\#\text{eQTLs in functional component of } R_j}.$$

Note that the number of eQTLs in GTEx tissue G_i is a weighted count, with an eQTL weighted by the inverse of the number of GTEx tissues in which the variant is eQTL, such that $\sum_i p_{G_i|R_j} = 1$. This way eQTLs that are unique to tissue G_i are given higher weight relative to eQTLs that are shared across many tissues. For GTEx tissue G_i, to test whether there is an enrichment in the functional component of Roadmap tissue R_j, we compare

$p_{G_i|R_j}$ with

$$p_{G_i|R_{-j}} = \frac{\#\text{eQTLs in tissue } G_i \text{ in functional components excluding } R_j}{\#\text{eQTLs in functional components excluding } R_j}.$$

The null hypothesis is $H_0 : P_{G_i|R_j} = P_{G_i|R_{-j}}$ vs. $H_0 : P_{G_i|R_j} > P_{G_i|R_{-j}}$. We apply a two-sample proportion test for each Roadmap tissue R_j and report the Roadmap tissue with minimum p value in Table 4.3.

The eQTLs that we used in these analyses are all significantly associated SNP-gene pairs for eGenes in each of these 44 GTEx tissues, obtained using a permutation threshold-based approach as described by the GTEx Consortium The GTEx Consortium [2015] (see also https://
www.gtexportal.org/home/documentationPage#staticTextAnalysisMethods for more details).

C.1 LD score regression

The stratified LD score regression approach [Finucane *et al.*, 2015] uses two sets of SNPs, reference SNPs and regression SNPs. The regression SNPs are SNPs that are used in a regression of χ^2 statistics from GWAS studies against the "LD scores" of those regression SNPs. The LD score of a regression SNP is a numeric score which captures the amount of genetic variation tagged by the SNP. Here, following Finucane *et al.* [2015] we use as regression SNPs HapMap3 SNPs, chosen for their high imputation quality, and as reference SNPs those SNPs with minor allele count greater than 5 in the 379 European samples from the 1000 Genomes Project. We first compute tissue-specific scores using each of our methods for the $9,254,335$ SNPs with minor allele count greater than 5 in the 379 European samples from the 1000 Genomes Project, which we will subsequently use as our "reference SNPs" for LD score regression. In the stratified LD score regression approach, a linear model is used to model a quantitative phenotype y_i for an individual i:

$$y_i = \sum_{j \in G} X_{ij}\beta_j + \epsilon_i.$$

Here G is some set of SNPs, X_{ij} is the standardized genotype of individual i at SNP j, β_j is the effect size of SNP j, and ϵ_i is mean-zero noise. In this framework, $\boldsymbol{\beta}$, the vector of all the β_j, is modeled as a mean-0 random vector with independent entries, and the

variance of β_j depends on the functional categories included in the model. We have a set of functional categories $C_1, \ldots, C_C$, and the variance of a SNP's effect size will depend on which functional categories it belongs to:

$$\mathrm{Var}(\beta_j) = \sum_{c:j \in C_c} \tau_c.$$

Here τ_c is the per-SNP contribution to heritability of SNPs in category C_c. In Finucane *et al.* [2015], the authors show that under this model τ_c can be estimated through the following equation:

$$E[\chi_j^2] = N \sum_c \tau_c l(j,c) + 1.$$

Here χ_j^2 is the chi-squared statistic for SNP j from a GWAS study, N is the sample size from that study, and $l(j,c)$ is the LD score of SNP j with respect to category C_c, $l(j,c) = \sum_{k \in C_c} r_{jk}^2$. This equation therefore allows for the estimation of the τ_c via the regression of the chi-squared statistics from a GWAS study on the LD scores of the regression SNPs.

Here, we extend the stratified LD score by allowing SNPs to be assigned to a category C_c probabilistically, that is, we assume a probability p_{kc} that SNP k belongs to category C_c, and therefore that the variance of its effect size is affected by its membership in that category. This only involves minor changes to the above equations, namely, we have that

$$\mathrm{Var}(\beta_j) = \sum_{c:j \in C_c} p_{jc}\tau_c,$$

where p_{jc} is the probability that SNP j belongs to category C_c, and as above

$$E[\chi_j^2] = N \sum_c \tau_c l(j,c) + 1,$$

although now $l(j,c) = \sum_{k \in C_c} p_{kc} r_{jk}^2$, p_{kc} being the probability that SNP k belongs to category C_c. We can therefore still estimate the τ_c via the regression of the chi-squared statistics from a GWAS study on the LD scores of the regression SNPs, but in calculating these LD scores we weight the squared correlation of a SNP k with a regression SNP j by the probability that SNP k belongs to a particular category.

For each tissue and phenotype, and each of our functional scores, we fit a separate LD score regression model, including the LD score derived using the posterior probability that

each regression SNP is in the functional component in that tissue, to estimate the per-SNP contribution of SNPs that belong to that component to heritability. To control for overlap of the tissue-specific functional score with other functional categories, we use the same 54 baseline categories used in Finucane *et al.* [2015], which represent various non-tissue-specific annotations, including histone modification measurements combined across tissues, measurements of open chromatin, and super enhancers.